Screening for Mr. Good Bob: Another 'Starting Small' Paper

David Kreps
April 2018[*]

1. Introduction

When setting out in a potentially long-run relationship, the parties involved must each be concerned with the "intentions" of their prospective trading partner. A simple and stylized model of this situation concerns one party, Alice, who at dates $t = 0, 1, \ldots$ must decide whether to trust a (prospective) trading partner, Bob. If Alice trusts Bob, Bob can respond by either treating Alice well or poorly. Alice would happily engage with Bob if she knows he will trust her well, but she is uncertain about Bob's incentives. Bob might be the sort of person who prefers to treat Alice well, but he might also be someone who prefers not to be engaged by Alice and, worse still, he might be the sort of person who wishes to be engaged by Alice *so that* he can treat her poorly. Alice's problem, then, is to take steps that, as efficiently (for her as possible), identifies whether Bob is a good prospective partner, a bad prospective partner, or someone so evil that he wants to treat her poorly.

One obvious screening device is to "start small." Alice (if she can) begins by trusting Bob a little bit, so if he treats her poorly, it costs her little. As time passes, Alice increases her scale of engagement with Bob, in the (presumed hope) that Bob, if bad or evil, will screen himself "out."

This idea has been studied by Sobel (1985) and especially by Watson (1999a, 1999b). Watson (1999a) is the most germane to the model studied here: He studies a continuous time model in which Alice is uncertain about Bob's intentions and, at the same time, Bob is uncertain about what are Alice's intentions. The model here bears a lot of similarities to Watson's, with the following differences: (a) This model involves only one-sided uncertainty, which makes it easier to understand as an application of (more-or-less) classic screening theory. (b) This model uses discrete time (not a major difference). (c) In Watson's analysis, Alice and Bob are presumed to engage

[*] Do not quote or cite. For exploratory and discussion purposes only.

in an (unmodeled) negotiation process how their relationship will involve. Here, in the spirit of a portion of the market signaling literature, Alice takes the lead in setting the "terms of trade." (d) Most importantly, in Watson's models (and Sobel's before his), the "evil" type comes in one flavor, so (for instance) in Watson's model, all the action happens right at the start or (only) when engagements reach full scale. The simpler basic setting here allows me to consider how Alice will screen out evil types when there is more than one degree of evilness. We see in particular how this complicates Alice's problem, as well as how different possible levels of evilness interact in terms of the (equilibrium) payoffs that each receives.

An important aspect of the model studied here is that Alice is modeled as a Stackelberg leader *with strong powers of commitment.* While some justification for this modeling assumption can (and will) be offered, since time is of the essence in Alice's screening methods, it raises important questions of the sort that arise in the literature on the Coase conjecture (e.g., Gul, Sonnenschein, and Wilson, 1986) and more generally on "screening through time" (Swinkels, 1999): After some screening has taken place, Alice would like to "reset" how she will behave. (Alice's ability to commit is reminiscent of Watson (1999a); compare with Watson (1999b), where Alice and Bob can renegotiate their arrangement as time passes.) I'll comment further on this near the end of this paper, but I don't do much analysis of it.

2. *The basic model*

Consider two parties, Alice and Bob, engaged in an infinitely repeated game, with stages at times $t = 0, 1, \ldots$.

At each time t, Alice must decide (for now) on whether to engage with Bob or not. If she doesn't engage with him, her payoff for this period is 0. If she does engage with him, he can either treat her well or poorly. If he treats her well, her payoff is 1. If he treats her poorly, her payoff is $-A$ for some parameter $A > 0$. Bob's payoffs from the three possible outcomes emain unspecified for now. But both Alice and Bob seek to maximize the expectation of the discounted sum of their stage-game payoffs, with common discount factor δ.

This game form is the basis of two canonical models employed in the study of credibility and reputation: The first model is the Threat Game, depicted in Figure 1a; the second is the Promise Game, depicted in Figure 1b. In the Threat Game, Bob prefers that Alice not engage with him to treating her well; in the Promise Game, his preferences between these two outcomes are reversed. His payoff for the third outcome, where she engages him and he treats her poorly, is a second parameter, B. In the Threat Game, $B < 0$, so if engaged, Bob's short-run interests are to treat Alice well. In the Promise Game, $B > 1$; Bob's short-run interests are to treat Alice poorly. Hence, in the Threat Game, Bob threatens Alice that, notwithstanding his short-run interests, he will treat her poorly, so that (if she deems this a credible threat) she does not engage him. And in the Promise Game, Bob promises not to treat her poorly, so that (if the promise is credible) she will engage with him. The literature constructs equilibria in which the threat/promise is credible, in some papers with an infinite horizon and in others with a finite horizon but some uncertainty in Alice's mind whether Bob is a crazy/nice type.

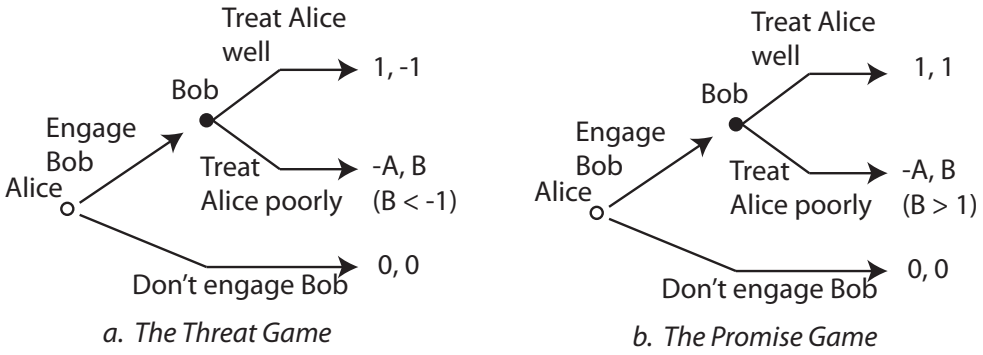

Figure 1. The Threat Game and the Promise Game. In both panels, Alice's payoffs are listed first, and $A > 0$. In panel a, $B < 0$; in panel b, $B > 1$.

We imagine that Alice faces Bob in the (infinitely) repeated game, where at the outset she doesn't know what are his preferences. She believes that he attaches some payoff to the three possible outcomes in the stage game, and that he seeks to maximize his discounted sum of stage-game paoyffs. But she (only) has a Bayesian prior over what his stage-game payoffs might be.

In terms of Bob's possible ordinal preferences over the three outcomes, and ruling out cases where there are ties, we group Bob into one of four

broad categories or types. In the first two types, Bob prefers treating Alice well to not engaging wih her. Normalize Bob's payoffs so that treating Alice well provides payoff 1 and nonengagement provides 0. The two types are:

G1. *Saintly Bob*, if his payoff from treating her poorly is $B < 1$

G2. *Good Bob*, if his payoff from treating her poorly is $B > 1$ (Note that in the Promise Game, Bob is Good Bob.)

In the second two types, Bob prefers non-engagement (payoff normalized to 0) to treating Alice well (payoff -1):

H1. *Bad Bob*, if his payoff from treating her poorly is $B < 0$. Bob from the Threat Game is Bad Bob with the stronger restriction that $B < -1$

H2. *Evil Bob*, if his payoff from treating her poorly is $B > 0$

The four (ordinal) types of Bob are depicted in Figure 2.

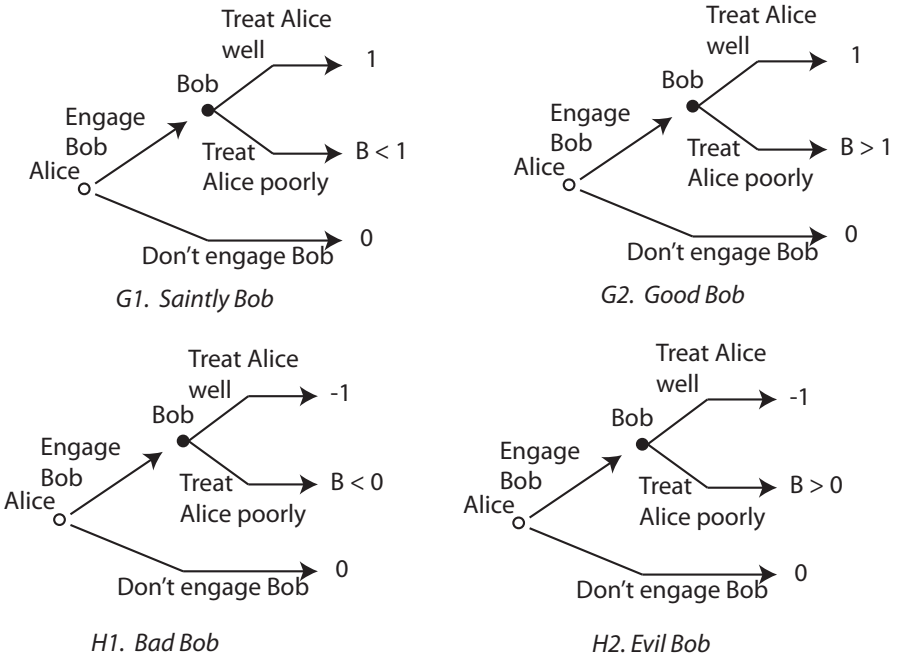

Figure 2. Four (ordinal) types of Bob

Alice's prior assessment concerning Bob's payoffs is given as follows: π_G between 0 and 1, is the probability that she faces a type of Bob who prefers to treat her well to not being engaged—that is, either Saintly Bob or

Good Bob—with $\pi_H = 1 - \pi_G$ the complementary probability that he is one of the two H types; F_G is the conditional cumulative distribution function for the parameter B, conditional on Bob being saintly or good; and F_H is the conditional cumulative distribution function for B, conditional on Bob being bad or evil. *I assume that F_H has finite support on $(0, \infty)$.*[1]

Because we are looking at a supergame formulation (with the added complication that Alice does not know Bob's "type"), many perfect (sequential or perfect-Bayes) equilibria can be constructed. I am interested in equilibria, if they exist, in which Alice's strategy takes a particularly simple and intuitive form:

> *At time 0, Alice engages with Bob. At any subsequent time (and starting from any other point in the game tree), if Bob has never in the past treated Alice poorly, she engages with him. But if at any previous time he treated her poorly, she does not engage with Bob.*

I offer no "formal" justification for limiting Alice's strategy in this way; it is certainly possible, for certain parameterizations, to construct equililbria that are Pareto superior to an equilibrium where Alice behaves in this fashion. But I believe, and hope to convince the reader, that looking for equilibria in which Alice employs this strategy (in a sense to be qualified subsequently) leads to some interesting economics.

What happens, then, if Bob understands that this is how Alice behaves?

- If Bob is saintly, he will treat her well. Doing so is provides him with his best stage-game payoff in all stages.
- Good Bob behaves precisely as he does in the standard stories about the Promise Game. If $B > 1/(1-\delta)$, he prefers treating her poorly once and getting 0's forever after to always treating her well, so he treats her poorly. If $B < 1/(1-\delta)$, he prefers to treat her well. (If $B = 1/(1-\delta)$, he has lots of best responses; to keep the story simple, I'll assume that F_G is continuous at the critical value $B = 1/(1-\delta)$, so this has zero prior

[1] This assumption is not entirely innocuous, as it entails the assumption that among the possible B's for Evil Bob, there is a least B greater than zero and a greatest B. The weaker assumption that the support of F_H on $(0, \infty)$ is contained within a compact set is sufficient for some of the formal results to come; the finiteness of the support is for expositional convenience.

probability.)

- Bad Bob will treat Alice poorly, unless the cost of doing so is so high that he prefers to treat her well forever. Simple calculations show that Bad Bob treats Alice poorly if $B > -1/(1-\delta)$ and well if the reverse inequality holds. (For simplicity, we assume that F_H is continuous at the critical value $B = -1/(1-\delta)$.) This implies that Bad Bob treats Alice poorly for all B satisfying $0 \geq B > -1$; it is only values of B that are less than -1 that might cause him to treat her well, which is, of course, the story in the Threat Game.
- Evil Bob will treat her badly. His best outcome is treating her badly, but given the assumption about her behavior, he will only ever have one shot at doing so. Since payoffs are discounted, he strictly prefers to take that one shot as soon as possible and then, subsequently, get his second favorite outcome every time.

So, if we suppose that Alice follows the strategy given above, beginning at time 0 by engaging Bob, the probability that she will be treated well (parameterized by δ is

$$\phi_\delta = \pi_G \, F_G\left(\frac{1}{1-\delta}\right) + \pi_H \, F_H\left(-\frac{1}{1-\delta}\right).$$

(Note that $\lim_{\delta \to 0} \phi_\delta = \pi_G$.) Hence, her expected payoff from following the strategy posited for her is

$$\phi_\delta\left(\frac{1}{1-\delta}\right) - \left[1 - \phi_\delta\right] A. \tag{1}$$

This is (weakly) positive if

$$\phi_\delta\left(\frac{1}{1-\delta}\right) - \left[1 - \phi_\delta\right] A \geq 0 \quad \text{which is true if} \quad \phi_\delta \geq \frac{(1-\delta)A}{(1-\delta)A + 1}.$$

And in this case, we have an equilibrium in which Alice behaves as we have posited. Note that as $\delta \to 1$, the condition for this being an equilibrium is that $\pi_G > A/(A+1)$.

(Is this equilibrium perfect? It is, as long as there is positive prior probiability that Bob is evil. Without going into all the details, note that we framed Alice's strategy as "engage with Bob if Bob has never treated her poorly in the past. This means that any errors by Alice do not affect her decision rule. If, for instance, she engages at time 0, is treated well, and then fails to engage at time 1, it is as if the time 1 interaction did not happen in terms of the continuation game. The argument is a bit more complex if we reach a point where she was treated poorly in the past, but the presence of Evil Bob is then used.)

2. Screening Bad Bob (and some Good Bob's) with cheap talk

But Alice can do better than this. If inequality 1 fails to hold, so we don't (yet) have an equilibrium, Alice can undertake a slight change in the rules of the game that may "restore" an equilibrium of this type. And even if inequality 1 holds, the same change in the rules may (probably will) improve her payoff. The change is a bit of cheap talk right at the start, where Alice asks Bob,

Bob: Do you want me to engage with you?

Bob's response to this is relatively straightforward. Saintly Bob certainly wants Alice to engage, as does Good Bob, although Good Bob's motives depend on his B: If Good Bob's B is less than $1/(1-\delta)$, he wants engagement so he and Alice can have a long-term, mutually beneficial relationship; if his B exceeds $1/(1-\delta)$, he wants Alice to engage with him so he can treat her poorly once. Evil Bob also wants Alice to engage with him, fully intending to treat her poorly at the first opportunity. As for Bad Bob, his answer is unequivocal. He does not want Alice to engage with him, since if she does not engage, he immediately (and forever) gets his best payoff.

Does Alice profit if she asks this question and then follows the request of Bob? She probably does, especially for δ close to 1. She gains by immediately screening out all the Bad Bob types who were going to treat her poorly, which is all the Bad Bob's with B such that $0 \geq B \geq -1/(1-\delta)$. On the other hand, she lets off the hook those Bad Bobs with B such that $B < -1/(1-\delta)$,

who prefer not to engage Alice but for whom signaling this by treating her poorly is exhorbitant. One expects, at least for δ close to one, this is a good tradeoff for Alice to make. I'll assume so, which fulfills the promise made last paragraph; for some parameterizations for which inequality 1 fails to hold, eliminating the Bad Bob's with cheap talk may turn Alice's expected payoff from negative to positive. And even if inequality 1 holds, eliminating the Bad Bob's is likely to improve her expected payoff.

This, of course, is not the end of the story. After Alice engages in this cheap talk, she still has two types of Bob who intend to treat her poorly: Good Bobs with $B > 1/(1-\delta)$ and Evil Bobs. The presence of Good Bobs with such high values of B complicates the analysis, without changing the basic story. To keep the exposition simple, I'll assume henceforth that the support of F_G lies entirely in $(-\infty, 1/(1-\delta))$. And, having eliminated all the Bad Bobs, I'll assume henceforth that $\pi_H = 1-\pi_G$ is the prior probability that Bob is Evil Bob, so that F_H has (finite) support that is a subset of $(0, \infty)$. In particular, the support of F_H will be denoted $\{B_1, B_2, \ldots, B_N\}$, where 0 $< \mathrm{B}_1 < B_2 < \ldots < B_N$, and ϕ_n is the (conditional) probability that Bob, if evil, has paramenter B_n.

It should now be evident why Evil Bob is so named. This type of Bob prefers non-engagement to treating Alice well, but even more than non-engagement, he would like to take his one shot at treating her poorly. Alice's problem is not so much in screening out Bad Bob types. They can be gotten rid of simply by asking them if they want to engage with her. It is Evil Bob about whom she must worry.

3. Screening out Evil Bob by starting small

Evil Bob can be screened out by several different means. For instance, Alice could try to institute and enforce a liquidated damages contract in which Bob must pay her the equivalent of $1 + A$ whenever he treats her poorly. Depending on the relative sizes of A and B—more precisely, on the payoff impact on Evil Bob of making a transfer to Alice sufficient to make her whole, this could either immediately screen out Evil Bob, who refuses to sign the contract, or, if B is large relative to A, could open up the possibility

that Alice and Evil Bob find their way to an arrangement where Evil Bob treats here poorly in every period by compensates her sufficiently so that she is happy.

I proceed in a different direction. Specifically, I will assume that Alice has the ability to "scale" her engagement with Bob, choosing at time t any scale ρ_t between 0 and 1: In the engagement at time t, if the scale of engagement is ρ_t, then all payoffs for Alice and for all types of Bob are ρ_t times their "full-engagement" values.

It is at this point that the normalizations of Bob's payoffs that were chosen become important. In particular, the normalization of Evil Bob's payoffs bear scrutiny. In this normalization, if Alice determines not to engage Evil Bob, his payoff is unaffected. But if she engages him and he treats her well, the cost he bears is proportional to the scale of the engagement, as is the benefit he receives from treating her poorly. It is reasonable to think that his cost and benefit will be monotonic functions of the scale of the engagement, but we are assuming much more, namely that they decrease in the same proportions. And even more than that, the same proportionality applies to Alice's payoffs from engaging him and either being treated well or poorly. (For reasons that will become clear, the assumptions on this point about the payoffs of the type-G Bobs are not of consequence.) This particular parameterization will make the algebra relatively simple. But it might be of interest to see what happens if, say, his benefit from treating her poorly reduces at a slower rate then his cost of treating her well (say, because part of his benefit is sadic pleasure he derives), or vice versa, or if her payoffs on either side scale at different rates, or if they scale at different rates than do his.

The assumption is that the choice of each ρ_t is Alice's choice to make. This presents us with two possibilities: Alice might be able to commit at the outset to the scales of engagement she will use. Or she might only be able to pick ρ_t after the time $t-1$ encounter is over. I deal here with the simpler case where Alice can and must commit. At the outset, she makes a (somehow) credible commitment of the form:

> *Bob: As long as you treat me well, at time t I will engage you at level ρ_t for the following sequence of engagement rates, $\{\rho_0, \rho_1, \ldots\}$. If you ever treat me badly,*

subsequently, I will not engage you at all (or, equivalently, ρ_t subsequently will be replaced by 0).

Skepticism concerning Alice's ability to make such a commitment is certainly warranted. It does not work to suppose that Alice can sign a binding contract with Bob to behave in this fashion, because (as we'll see) events can transpire that would make it in the interests of Alice and Bob to rip such a contract. A better story, perhaps, is that Alice has this encounter with a many Bobs through time—she is an employer, employing many Bobs, say—and she wishes to protect a reputation for keeping to any such announcement. Still, the reader is entitled to be skeptical.

Skepticism aside, suppose Alice can commit in this fashion. Is there any point in using it? The idea, at least at first blush, is to get Evil Bob to reveal himself at a lower and therefore less costly scale. (A second way that starting small can help Alice will be developed.) Of course, if Alice announces $\{\rho_t; t = 0, 1, \ldots\}$ (and is committed to this sequence), Evil Bob will choose the time t to treat Alice poorly that is best from his perspective. Alice can trust Saintly Bob to always treat her well; Good Bob, on the other hand, might find it worthwhile to treat Alice poorly, depending on $\{\rho_t\}$. Her problem, then, is to make a commitment to a sequence of scales $\{\rho_t\}$ that provides her with the greatest possible expected value, given that Good Bob and Evil Bob will react to this commitment as Stackelberg followers.

When Evil Bob is indifferent among several different times for treating Alice poorly, we make the usual assumption that he chooses among those times the time that is best for Alice.

Alice faces the following basic trade-off. Conditional on facing Evil Bob, she likes low values of ρ_t. But, if Bob is saintly or good (and assuming for the moment that Good Bob will always treat her well), she wants $\rho_t = 1$ or, barring that, to be as large as possible. Having small ρ_t for small t and bigger ρ_t for larger t—starting small—would seem the way to go. But Alice can't, for instance, have a very small ρ_0 and then set $\rho_t = 1$ for all subsequent t's; Evil Bob, hearing this, will wait until time 1. Alice's optimization problem, then, is to find the best way to compromise between these her two conflicting desires.

4. Some preliminary analysis

To avoid repeatedly using the phrase "treating Alice poorly," I substitute "triggering" to describe Evil Bob taking his one shot at treating Alice poorly.

The following Lemma will be used repeatedly.

Lemma 1. *Suppose Alice announces $\{\rho_t\}$. Suppose Evil Bob, with parameter B, is considering whether it is better to trigger at time t or $t+1$. (This is not to say that one of those two is optimal for him. This only considers which is better for him.) If $\rho_t/\rho_{t+1} < \delta B/(B+1)$, he strictly prefers to trigger at $t+1$. If $\rho_t/\rho_{t+1} > \delta B/(B+1)$, he strictly prefers to trigger at time t. If ρ_t/ρ_{t+1}, he is indifferent.*

Hence, if for two times t' and t'', with $t'' > t'$, Alice's announcement satisfies

$$\rho_t = \rho'_t \left[\frac{\delta B}{B+1}\right]^{t''-t},$$

$t = t', t'+1, \ldots, t''$, Evil Bob with parameter B is indifferent among triggering at times $t', t'+1, \ldots, t''$; Evil Bob with parameter $B' < B$ strictly prefers to trigger at time t' to any time $t'+1, \ldots, t''$; and Evil Bob with parameter $B' > B$ strictly prefers to trigger at time t'' to any time $t', t'+1, \ldots, t''-1$.

Proof. Fixing $\{\rho_s\}$, the difference between Bob's payoff from triggering at times t and triggering at time $t+1$ is

$$\left[\delta^t \rho_t B - \sum_{s=0}^{t-1} \delta^s \rho_s\right] - \left[\delta^{t+1} \rho_{t+1} B - \sum_{s=1}^{t} \delta^s \rho_s\right] = \left[\delta^t \rho_t - \delta^{t+1} \rho_{t+1}\right] B + \delta^t \rho_t.$$

The sign of this difference is the sign of this term divided by δ^t, which is

$$\left[\rho_t - \delta \rho_{t+1}\right] B + \rho_t.$$

If $\rho_{t+1} = 0$, Evil Bob certainly prefers to trigger at time t over time $t+1$ (unless ρ_t also equals zero; if $\rho_t = \rho_{t+1} = 0$, he is indifferent), and if $\rho_{t+1} > 0$, the sign of the differentce is the same as the sign of

$$\left[\frac{\rho_t}{\rho_{t+1}} - \delta\right] B + \frac{\rho_t}{\rho_{t+1}}.$$

This strictly exceeds zero if and only if

$$\frac{\rho_t}{\rho_{t+1}}(B+1) > \delta B \quad \text{or} \quad \frac{\rho_t}{\rho_{t+1}} > \frac{\delta B}{B+1}.$$

The reverse inequality and equation follow similarly.

The second paragraph in the Lemma follows immediately from iterated application of the first paragraph. ∎

Proposition 1.

a. Suppose Alice announces $\{\rho_t\}$ with at least one ρ_t strictly positive. Then Evil Bob with any parameter B has a finite solution to his optimization problem of when to trigger. In fact, setting $\overline{\rho} := sup\{\rho_t; t = 0, 1, \ldots\}$ and letting τ be the earliest time such that $\rho_t > \delta\overline{\rho}$, Evil Bob (regardless of his B) will trigger at time τ or earlier.

b. Suppose that Alice announces $\{\rho_t\}$ with $\rho_t > 0$ for some t. Suppose Evil Bob with parameter B has, as an optimal solution, to trigger at t (not precluding the possibility that he has several solutions). Then: (i) for $B' < B$, any and all optimal triggering times solutions for Evil Bob with parameter B' are to trigger at time $t' \leq t$. (ii) For $B'' > B$, any and all optimal triggering times for Evil Bob with parameter B'' are to trigger at time t or later.[2] *(iii) Good Bob with parameter $\hat{B}$, for any $\hat{B}$, will never treat Alice poorly before time t.*

c. If the sequence $\{\rho_t\}$ is non-decreasing, Good Bob will always treat Alice well.

d. As long as $\pi_G > 0$, Alice can obain a strictly positive payoff in this game.

e. Alice's optimal announcement (any one, if there are ties) has $\rho_0 > 0$ and $\rho_t = 1$ for all $t > \tau$, for τ as defined in part a.

The possibility that Evil Bob with a particular value of B is indifferent between triggering at several different times t is an important property of Alice's optimal announcement. When this happens—that is, whenever Evil

[2] The second conclusion is a trivial repackaging of the first, but is included since this property is used extensively in both forms. From the perspective of the general theory of screening, this is the usual "single-crossing" property.

Bob with parameter B has more than one optimal triggering time—unless $A = B$, Alice cares which of these times Bob chooses. Our assumption that all stage-game payoffs scale linearly in ρ makes it simple to say what Alice prefers:

Proposition 2. *Suppose Alice announces $\{\rho_s\}$. Suppose that Evil Bob with parameter B is indifferent between triggering at times t and t', for $t > t'$ and such that $\rho_{t'} > 0$. If $A > B$, Alice strictly prefers that Bob trigger at time t'. If $A < B$, she strictly prefers that he choose t. If $A = B$, she is indifferent.*

Proof of Proposition 1. (a) Suppose $\{\rho_t\}$ is Alice's announcement, $\overline{\rho} = \sup\{\rho_t : t = 0, 1, \ldots\}$, and τ is the earliest time at which $\rho_t > \delta\overline{\rho}$. The difference between Evil Bob's payoff (with any parameter B) from triggering at time τ versus triggering at any time $t > \tau$ is

$$\delta^\tau \rho_\tau B - \delta^t \rho_t B + \sum_{s=\tau}^{t-1} \delta^s \rho_s = \left[\delta^\tau \rho_\tau - \delta^t \rho_t\right] B + \sum_{s=\tau}^{t-1} \delta^s \rho_s.$$

From the way τ is defined, it is evident that $\rho_\tau > 0$, so the summation term on the right-hand side of the last display is strictly positive. And the first term is also strictly positive, since $\rho_\tau > \delta\rho_t$ and, inside the square brackets, ρ_t is discounted by $t - \tau$ more δ's than is ρ_τ. Since this puts a finite upper bound, uniform in B, on when Evil Bob will trigger, he is optimizing over finitely many terms, so of course his problem has a solution.

Although not mentioned in the statement of the proposition, the same result holds for Good Bob, with one amendment: Evil Bob will certainly trigger at some point; we now know that this is at or before time τ. If Good Bob triggers, it will be at or before τ, but Good Bob may never trigger; see part c.

(b) If Alice announces $\{\rho_t\}$ with some strictly positive ρ_t, Evil Bob's optimal solution gives him a strictly positive payoff, because he can always trigger at the first time t that $\rho_t > 0$. Suppose that t is an optimal triggering time for Evil Bob with parameter B. Then triggering at t must be at least as good

as triggering at any time $t' > t$, which is

$$\delta^t \rho_t B - \sum_{s=0}^{t-1} \delta^s \rho_s \geq \delta^{t'} \rho_{t'} B - \sum_{s=0}^{t'-1} \delta^s \rho_s,$$

which can be rewritten

$$\left[\delta^t \rho_t - \delta^{t'} \rho_{t'}\right] B \geq - \sum_{s=t}^{t'-1} \delta^s \rho_s.$$

Of course, for t to be optimal, it must be that $\rho_t > 0$, so the right-hand side of the previous inequality is strictly negative. This implies that, for $B' < B$,

$$\left[\delta^t \rho_t - \delta^{t'} \rho_{t'}\right] B' > - \sum_{s=t}^{t'-1} \delta^s \rho_s;$$

if $\delta^t \rho_t - \delta^{t'} \rho_{t'} \geq 0$, then multiplying by $B' > 0$ leaves a nonnegative term which is strictly greater than the negative term on the r.h.s., while if $\delta^t \rho_t - \delta^{t'} \rho_{t'} < 0$, then $\left[\delta^t \rho_t - \delta^{t'} \rho_{t'}\right] B' > \left[\delta^t \rho_t - \delta^{t'} \rho_{t'}\right] B$, and we have the desired strict inequality. In either case, Evil Bob with parameter $B' < B$ strictly prefers triggering at t to any subsequent time, and so his optimal time to trigger must be t or less.

Reiterating from fn.3, (ii) follows immediately from (i). If $B' > B$, t is optimal for Evil Bob with parameter B, and t' is optimal for Evil Bob with parameter B', then we know that $t \leq t'$, which of course is the same as $t' \geq t$.

And if t is optimal for Evil Bob with parameter B, then triggering at $t' < t$ is no better for him than triggering at t, or

$$\delta^t \rho_t B - \sum_{t'}^{t-1} \delta^s \rho_s \geq \delta^{t'} \rho_{t'} B.$$

If $\rho_{t'} = 0$, then since Evil Bob can get a strictly positive payoff, his optimal payoff must be strictly positive, and we get a strict inequality in the last

display. And if $\rho_{t'} > 0$, then the summation in the last display is strictly positive. In either case, we have

$$\text{hence} \quad \delta^t \rho_t B > \delta^{t'} \rho_{t'} B,$$

and so for any $\hat{B} > 1$, $\delta^t \rho_t \hat{B} > \delta^{t'} \rho_{t'} \hat{B}$. That is, Good Bob with any $\hat{B}$ prefers strictly to wait for t to trigger, if he will trigger at all, without even taking into account the positive flow of payoffs he gets from waiting. That gives (iii).

(c) Suppose $\{\rho_t\}$ is non-decreasing. If Bob is good (with parameter B), and if he hasn't triggered prior to time t, his continuation payoff (in present-value terms) starting at t if he always treats Alice well is $\sum_{s=t}^{\infty} \delta^{s-t} \rho_s \geq \sum_{s=t}^{\infty} \delta^{s-t} \rho_t = \rho_t/(1-\delta)$. If he triggers at time t, his payoff (in present-value terms) is $\rho_t B$. We assumed that the support of B for type-G Bobs did not extend to $B > 1/(1-\delta)$, so always treating Alice well is unimprovable, hence optimal, for Good Bob.

(d) Let $\overline{B}$ be the largest value of B for Evil Bob, plus 1. Let $\gamma = \delta\overline{B}/(\overline{B}+1)$. Suppose Alice announces $\{\rho_t$ given by $\rho_t = 1$ for $t \geq T$ and $\rho_t = \gamma^{T-t}$ for $t < T$, for some (presumably large) T. Then for all t, $\rho_t/\rho_{t+1} \leq \gamma = \delta\overline{B}/(\overline{B}+1) > \delta B/(B+1)$ for all possible values of Evil Bob's B. Hence, by Lemma 1, all Evil Bobs (that is, whatever is Evil Bob's parameter B) will trigger at time 0. And since $\{\rho_t\}$ is nondecreasing, all type-G Bobs will treat Alice well at all dates. Hence Alice's expected payoff from this announcement is

$$\pi_G \left[\sum_{s=0}^{T-1} \delta^s \gamma^{T-s} + \frac{\delta^T}{1-\delta} \right] - \pi_H \gamma^T A.$$

To explain, the term pre-multiplied by π_G is the sum of payoffs she receives if Bob is type G; if Bob is evil, she immediately loses A *scaled down by scale factor* γ^T. Even ignoring the summation term (which is positive), since $\delta > \gamma = \delta\overline{B}/(\overline{B}+1)$, the term $\pi_G \delta^T/(1-\delta)$ goes to zero more slowly than does $\pi_H \gamma^T A$ (as long as $\pi_G > 0$), and so for large enough T, Alice has a strictly positive payoff.

(e) Since Alice can obtain a strictly positive payoff, we know that her optimal announcement $\{\rho_t\}$ must have some $\rho_t > 0$. Suppose that, in this announcement, $\rho_0 = 0$. Let t^* be the first time that $\rho_t > 0$, and consider what happens if Alice instead announces $\{\rho'_t\}$ given by $\rho'_t = \rho_{t+t^*}$. Bob (of any stripe), as a Stackelberg follower (who breaks ties in whatever way favors Alice), will do whatever he would have done against $\{\rho_t\}$, except t^* periods earlier. This increases Alice's expected payoff by $1/\delta^{t^*}$, because of less discounting. So if $\rho_0 = 0$, $\{\rho_t\}$ cannot be optimal for Alice.

Again begin by assuming that $\{\rho_t\}$ is optimal for Alice, and let $\overline{\rho} := \sup\{\rho_t; t = 0, 1, \ldots\}$, and τ be the first time that $\rho_\tau > \delta\overline{\rho}$. From part a, we know that Evil Bob will never (optimally) trigger after time τ. If we replace ρ_t for $t > \tau$ with $\overline{\rho}$, this doesn't change; Evil Bob still prefers to trigger at time τ to any later time. Hence this change has no effect on Alice's payoff conditional on Bob being evil; and it can only improve her payoff if Bob is type-G. (If Bob is good and has already triggered, this change can only induce him to trigger later or not at all, both of which are good for Alice. If Bob is good and has not already triggered, or if he is saintly and so has certainly not triggered, this clearly is good for Alice.) Now suppose that $\overline{\rho} < 1$. Replace every ρ_t with ρ'_t where $\rho'_t = \rho_t/\overline{\rho}$. Bob makes the same choices he did before. And this changes Alice's *overall* expected payoff by a factor of $1/\overline{\rho} > 1$. Since we know her overall expected payoff is strictly positive, this raises her overall payoff. ∎

Proof of Proposition 2. If Evil Bob with parameter B is indifferent between triggering at times t and t', with $t > t'$, then

$$\left[\delta^t \rho_t - \delta^{t'} \rho_{t'}\right] B = \sum_{s=t'}^{t-1} \delta^s \rho_s.$$

Since $\rho_{t'} > 0$, the right-hand side of the equation is strictly positive, so the left-hand side must be as well, hence $\delta^t \rho_t - \delta^{t'} \rho_{t'} > 0$. But then if $A > B$,

$$\left[\delta^t \rho_t - \delta^{t'} \rho_{t'}\right] A > \left[\delta^t \rho_t - \delta^{t'} \rho_{t'}\right] B = \sum_{s=t'}^{t-1} \delta^s \rho_s, \quad \text{and therefore}$$

$$\sum_{s=t'}^{t-1} \delta^s \rho_s - \delta^t \rho_t A < -\delta^{t'} \rho_{t'} A \quad \text{or} \quad \sum_{s=0}^{t-1} \delta^s \rho_s - \delta^t \rho_t A < \sum_{s=0}^{t'-1} \delta^s \rho_s - \delta^{t'} \rho_{t'} A.$$

This says that if $A > B$, conditional on Bob being Evil Bob (with parameter B), Alice prefers him to trigger sooner rather than later. Her payoffs from Bob conditional on Bob being of type G are unaffected by Evil Bob's choice so, overall, she wants Bob to trigger sooner. Of course, if $B > A$, the reverse inequality will hold. ∎

Please note: In Proposition 2, the antecendent is that Evil Bob with parameter B is indifferent between triggering at t and t' and *not necessarily* that these are optimal triggering times for him. The proposition could alternatively be phrased, "If Alice announces $\{\rho_s\}$ with some $\rho_s > 0$, and t and t' are both optimal for Evil Bob with parameter B, then…" This follows becasue if some $\rho_s > 0$, then $\rho_t > 0$ must be true for any optimal triggering time for Evil Bob.

5. The form of Alice's optimal announcement

Recall the assumption that the support of Evil Bob's parameters B is a finite set, $\{B_n; n = 1, \ldots, N\}$, where we assume these are ennumerated in ascending order: $0 < B_1 < B_2 < \ldots < B_N$. For $n = 1, \ldots, N$, let $\gamma_n = \delta B_n/(B_n + 1)$. And, henceforth, we refer to Evil Bob with parameter B_n as B_n-EB.

Proposition 3. *An optimal announcement for Alice, $\{\rho_t\}$, takes the following form: For a set of times $0 = \tau_0 \leq \tau_2 \leq \ldots \leq \tau_N$, ρ_t is as follows:*

$$\rho_t = \begin{cases} 1, & \text{for } t \geq \tau_N, \text{ and} \\ \\ \rho_{t_n} (\gamma_n)^{\tau_n - t}, & \text{for } t \text{ between } \tau_{n-1} \text{ and } \tau_n. \end{cases}$$

This sequence $\{\rho_t\}$ causes B_n-EB to be indifferent among triggering at any time from τ_{n-1} to τ_n. And hence, on Alice's behalf, B_n-EB triggers at time t_{n-1} if $B_n < A$ and at time t_n if $B_n > A$.

Several remarks about this proposition are in order.

1. Most importantly, this characterizes Alice's optimal announcement, but it does not solve her problem. It (merely) reduces her problem to finding the best sequence of $\{\tau_n\}$, which is not a trivial problem, even if there is only one type of Evil Bob. (We'll discuss the full solutions for the case of one type of Evil Bob and two types in later sections.) And it characterizes one optimal announcement. In knife-edge cases, she may have more than one.

2. Note that $\tau_n \leq \tau_{n+1}$ is allowed. Imagine, for instance, that $N = 6$, $B_6 < A$, and $0 = \tau_0 = \tau_1 < \tau_2 < \tau_3 = \tau_4 = \tau_5 < \tau_6$. Since $B_6 < A$, $B_n < A$ for all n, and all types of Evil Bob trigger at the start of "their" interval: Types B_1 and B_2 both trigger at $t = 0$, type B_3 triggers at τ_2, and types B_4, B_5, and B_6 all trigger at $\tau_3 = \tau_4 = \tau_6$. Pooling is certainly a possibility.

3. The formula for ρ_t is "recursive" in the following sense. Given τ_N, we have $\rho_{\tau_N} = 1$, and so, for t between τ_{N-1} and τ_N, $\rho_t = \rho_{\tau_N} \cdot (\gamma_N)^{\tau_N - t} = 1 \cdot (\gamma_N)^{\tau_N - t}$. In particular, $\rho_{\tau_{N-1}} = (\gamma_N)^{\tau_N - \tau_{N-1}}$. This establishes the value of $\rho_{\tau_{N-1}}$, and then the definition allows us to compute ρ_t for t from τ_{N-2} to τ_{N-1}. And so forth.

4. Suppose $A = 15$, $B_2 = 10$, and $B_3 = 20$. (The point is that from B_2 to B_3, the B values jump from below A to above A.) Then, according to the proposition, B_2-EB will trigger at time τ_1, while B_3-EB will trigger at time τ_3. The sequence $\{rho_t\}$ is allowed to have a "break point" at an intermediate time τ_2, but that is only τ_n at which no type of Evil Bob triggers.

   A picture may clarify how this works. To make the picture clearer, I choose extreme values. Suppose there are three values of B, $B_1 = 3, B_2 = 10$ and $B_3 = 100$. If $\delta = 0.9$, we have $\gamma_1 = 0.675$, $\gamma_2 = 0.8182\ldots$, and $\gamma_3 = 0.89108910\ldots$ Suppose that Alice chooses $\tau_1 = 8, \tau_2 = 12$, and $\tau_3 = 15$ Figure 3 provides a graph of the ρ_t that results. Note that if $A = 7$ in this example, B_1-EB would trigger at $t = 0$, B_2-EB would trigger at $t = 12$, and B_3-EB would trigger at $t = 15$. If A were instead 20, B_2-EB would trigger at 8. But, for these numbers, B_1-EB is indifferent about

triggering any time between 0 and 8, inclusive; B_2-EB is indifferent for t between 8 and 12; and B_3-EB is indifferent among t between 12 and 15.

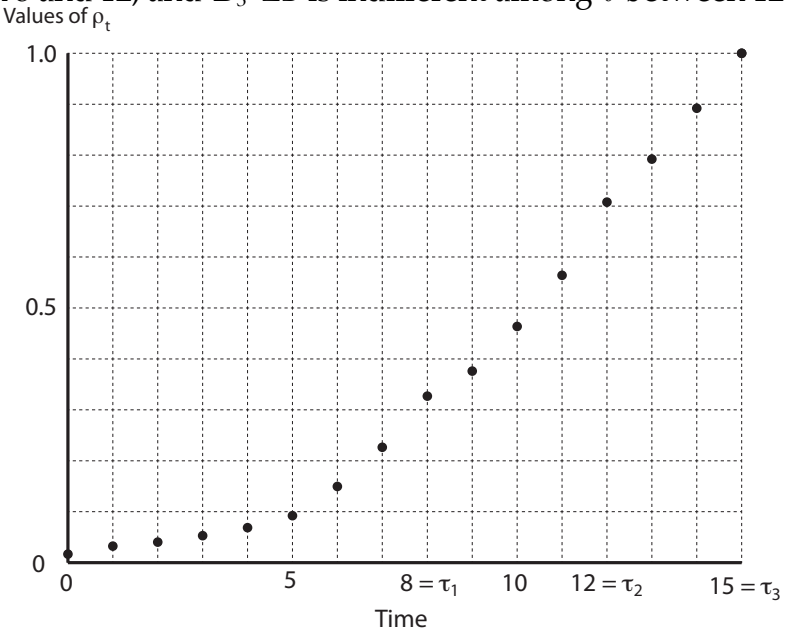

Figure 3. The shape of an "optimal announcement" by Alice. The ρ_t's "descend" geometrically from the value 1 at $\tau_3 = 15$ until $\tau_2 = 12$ at the rate γ_3, then they descend at the higher rate γ_2 until $\tau_1 = 8$, and then at rate γ_1 until $\tau_0 = 0$. Hence B_1-EB sees triggering at any time from τ_0 to τ_1 as optimal, B_2-EB sees times from τ_1 to τ_2 as optimal, and so forth..

5. From Lemma 1, it is clear that *if* Alice announces a sequence of scales $\{\rho_t\}$ that takes this form, then B_n-EB's optimal response is to trigger sometime between τ_{n-1} and τ_n, and he is indifferent among those times: Prior to time τ_{n-1}, the ratio $\rho_t/\rho_{t-1} = \gamma_{n'}$ for some $n' < n$. Since $n' < n$ implies $\gamma_{n'} > \gamma_n$, the scale factors are increasing at a rate faster than the rate that keeps B_n-EB indifferent, and so he wishes to wait. After time τ_n, the ratio is $\gamma_{n'}$ for some $n' > n$, hence at a rate that is "too slow" for B_n-EB. And between τ_{n-1} and τ_n, $\rho_t/\rho_{t+1} = \gamma_n$, which keeps B_n-EB indifferent. (If, say, $\tau_{n-1} = \tau_n$, the ratio ρ_t/ρ_{t+1} is "too fast" for B_n-EB prior to τ_n—he wishes to wait—and it is "too slow" after τ_n—he prefers to trigger right away, so his *unique* optimal response is to trigger at $\tau_{n-1} = \tau_n$.)

So if Alice makes an announcement of this form, Evil Bob will respond as described. Since the sequence $\{\rho_t\}$ is nondecreasing, Good Bob will always

treat Alice well. The hard work in proving the proposition is to show that this is the form of Alice's *optimal* announcement.

Proof. Suppose that $\{\rho_t\}$ is indeed the optimal announcement by Alice. We know from last section that $\rho_0 > 0$ and, for sufficiently large T, $\rho_t = 1$ for all $t > T$.

We also know that if we write $\mathcal{T}(n)$ for the set of times t that are optimal triggering dates for B_n-EB, then the sets $\mathcal{T}(n)$ are strongly ordered in n: If $t \in \mathcal{T}(n)$, then, for $n' < n < n''$, all $t' \in \mathcal{T}(n')$ are less than or equal to t and all $t'' \in \mathcal{T}(n'')$ are greater than or equal to t. For each n, let t_n^* be the best $t \in \mathcal{T}(n)$ for Alice; that is, if $B_n < A$, then t_n^* is the least element of $\mathcal{T}(n)$, while if $B_n > A$, t_n^* is the biggest element of $\mathcal{T}(n)$. If $B_n = A$ for some n, lump it in with the cases $B_n < A$; that is, let t_n^* be the least member of $\mathcal{T}(n)$. Of course, we immediately have that $t_1^* \leq t_2^* \leq \ldots \leq t_N^*$. Also, we know that if Good Bob is induced to treat Alice poorly by $\{\rho_t\}$, it can be no earlier than t_N^*.

Suppose we wish to maximize Alice's expected payoff *maintaining that, for each n, t_n^* is an optimal triggering time for B_n-EB.* Then for each t we have N "incentive" constraints, namely that t provides a payoff for B_n-EB no greater than does t_n^*. I assert that for each t, at least one of these constraints must be binding and, moreover, if $t_n^* < t < t_{n+1}^*$, the binding constraint(s) must include either the constraint that B_n-EB weakly prefers t_n^* to t or B_{n+1}-EB weakly prefers t_{n+1}^* to t. For suppose that t is such that $t_n^* < t < t_{n+1}^*$, B_n-EB strictly prefers t_n^* to t, and B_{n+1}-EB strictly prefers t_{n+1}^* to t. Then Alice can increase ρ_t by some small amount, small enough so that t_n^* remains optimal for B_n-EB and t_{n+1}^* remains optimal for B_{n+1}-EB. And this change has no impact on the triggering decisions of any other variety of Evil Bob: For $B_{n'}$-EB with $n' < n$, it cannot be that t becomes optimal, because $t > t_n^*$, and t_n^* remains optimal for B_n-EB. And this increase in ρ_t only increases for $B_{n'}$-EB the costs he faces by triggering at any other date, an increased cost he avoids by triggering at $t_{n'}^* \leq t_n^*$. And for $n' \geq n+1$, $B_{n'}$-EB cannot optimize by triggering at t, because t_{n+1}^* remains optimal for B_{n+1}-EB, so the only candidates for optimal times for $B_{n'}$-EB are at time t_{n+1}^* or later. The increase in ρ_t increases the costs faced by $B_{n'}$-EB, but it

increases the costs of triggering at times t^*_{n+1} and later by the same amount, so doesn't affect his optimal choice.

But for Alice, this slight increase raises her expected payoff: It raises what she gets from any type-G Bob or $B_{n'}$-EB for $n' > n+1$ at time t. (Were Good Alice to trigger, it has to come after t^*_{n+1}.)

By a similar argument, for $t < t^*_1$, the constraint for B_1-EB must bind: If it does not, ρ_t can be increased slightly without affecting the optimality of t^*_1 for B_1-EB. And the optimality of t^*_n for all other B_n-EBs is unaffected: It can't be optimal for them to trigger before t^*_1, and this variation, while it raises their costs, raises costs after time t equally. And, for Alice, this slight rise increases her payoffs.

For $t > t^*_N$, we use a slightly different argument (to handle the possibility of Good Bob being induced to trigger). First, I assert that for all $t > t^*_N$, $\rho_t \leq \min\{1, \rho_{t^*_N}(\gamma_N)^{t^*_N - t}$. For if not, there is a first time $t' > t^*_N$ that this inequality is violated. Compare B_N-EB triggering at time t^*_N with triggering at time t': The cost to him of waiting until time t' is less than it would be if he faced $\rho_s = \rho_{t^*_N}(\gamma_N)^{t^*_N - s}$ for $s = t^*_N$ up to time $t'-1$. And his immediate (time t') payoff is greater than it would be if, at that time, he faced $\rho_{t^*_N}(\gamma_N)^{t^*_N - t'}$. If he faced the alternative sequence of ρ_t's for $t > t^*_N$, he would be indifferent (Lemma 1), so he is strictly better off triggering at time t', a contradiction. It is the case that for all $t > t^*_N$, $\rho_t \leq \min\{1, \rho_{t^*_N}(\gamma_N)^{t^*_N - t}$.

But then suppose Alice replaced ρ_t for $t \geq t^*_N$ with $\min\{1, \rho_{t^*_N}(\gamma_N)^{t^*_N - t}\}$. This keeps t^*_N optimal for B_N-EB, and it doesn't affect the optimality of t^*_n for B_n-EB for any other n. It ensures that Good Bob will always treat her well and, to the extent that any one of the ρ_t's is increased, it increases her payoff from type-G Bobs. Since $\{\rho_t\}$ is meant to be optimal for Alice, this implies that ρ_t for $t > t^*N$ is indeed $\min\{1, \rho_{t^*_N}(\gamma_N)^{t^*_N - t}$. (And, supposing that $B_N > A$, it implies that $\rho_{t^*_N}/\gamma_N > 1$.)

Consider next values of ρ_t for $t < t^*_1$: We know that the constraint for B_1-EB must bind for these times, which is to say that B_1-EB is indifferent between triggering at t^*_1 and at any earlier time. By Lemma 1 (and working back from time t^*_1, this implies that $\rho_t = \rho_{t^*_1}(\gamma_1)^{t^*_1 - t}$ for all $t \leq t^*_1$. (And, supposing that $B_1 < A$, this implies that $t^*_1 = 0$.)

Consider ρ_t for t between t_n^* and t_{n+1}^*, for any n such that $t_n^* < t_{n+1}^*$. We showed that for each such t, either the constraint for B_n-EB or the constraint for B_{n+1}-EB must bind. In fact, we know more: Define τ_n as the last time that the constraint for B_n-EB binds. Then: (i) this constraint binds for all t between t_n^* and τ_n; the constraint for B_{n+1}-EB binds for all t between $\tau_n + 1$ and t_{n+1}^*; (iii) and it is only for τ_n that the constraints for *both* B_{n+1}-EB and B_n-EB can bind.

This follows from Proposition 1: To say that the B_n-EB constraint binds at τ_n is to say that B_n-EB is indifferent between triggering at t_n^* and at τ_n. But since t_n^* is optimal for B_n-EB, this would say that τ_n is also optimal for him. And then, no t such that $t < \tau_n$ can be optimal for B_{n+1}-EB. On the other hand, by the definition of τ_n, the constraint for B_n-EB does not bind for $t > \tau_n$. This leaves only τ_n: By definition, this is an optimal triggering time for B_n-EB, and it *might also* be optimal for B_{n+1}-EB.

Enlisting Lemma 1, this means that for t from t_n^* up to and including τ_n, $\rho_t = \rho_{t_n^*}(\gamma_n)^{t_n^* - t}$, and for t from t_{n+1}^* down to *and possibly including* τ_n, $\rho_t = \rho_{t_{n+1}^*}(\gamma_{n+1})^{t_{n+1}^* - t}$.

To simplify the notation and exposition, define $\psi_n(t) = \rho_{t_n^*}(\gamma_n)^{t_n^* - \tau_{n+1}}$ and $\psi_{n+1}(t) = \rho_{t_{n+1}^*}(\gamma_{n+1})^{t_{n+1}^* - t}$.

Then: We know that $\rho_{\tau_n} = \psi_n(\tau_n)$. *I assert that at* τ_n, *both constraints must bind*; that is, $\psi_n(\tau_n) = \psi_{n+1}(\tau_n) = \rho_{\tau_n}$. Suppose by way of contradiction that $\rho_{\tau_n} = \psi_n(\tau_n) \neq \psi_{n+1}(\tau_n)$. In this case, it must be that $\psi_{n+1}(\tau_n) > \rho_{\tau_n}$; were the reverse inequality true, B_{n+1}-EB would prefer triggering at τ_n to triggering at t_{n+1}^*. On the other hand, since $\rho_{\tau_n+1} = \psi_{n+1}(\tau_n + 1)$, it must be that $\psi_n(\tau_n + 1) > \psi_{n+1}(\tau_n + 1)$; if the two were equal, τ_n would be one time unit to the right (since both constraints bind there), and if we had a $<$ inequality, B_n-EB would prefer triggering at $\tau_n + 1$ to triggering at t_n^*.[3] Graphically, the situation must be as depicted in Figure 4, where the two curves are the functions ψ_n and ψ_{n+1}, and solid dots represent values of ρ_t.

But this situation is inconsistent with Alice optimizing except for a knife-edge case and, in that knife-edge case, we can adjust the ρ_ts so that τ_n is

[3] In other words, the function $\psi_n(t)$ is the "indifference curve" for B_n-EB that passes through $\rho_{t_n^*}$. And, if it wasn't already obvious, this should clarify why we have a classic situation of single-crossing.

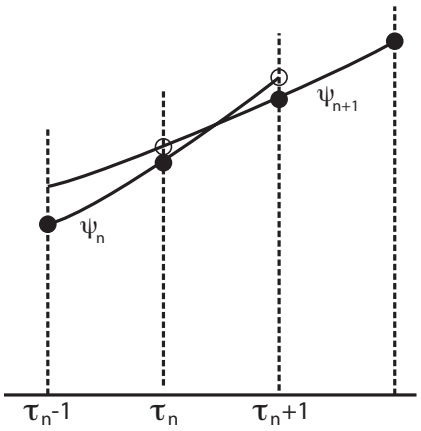

Figure 4. The hypothesized situation around τ_n. See text for explanation.

optimal for both B_n- and B_{n+1}-EB. To see this, let V_{τ_n} be Alice's expected value for payoffs she accrues from time 0 up to and including time τ_n. Suppose $V_{\tau_n} > 0$. Then, if Alice changes all her ρ_t from time 0 to time τ_n by multiplying each by $\lambda > 1$, while keeping λ below $\psi_{n+1}(\tau_n)/\psi_n(\tau_{n)}$, no decision by any Evil Bob changes: Those with $n' \leq n$ face the same (proportionally better) incentives, while those with $n' \geq n+1$ are not tempted to trigger earlier, because B_{n+1}-EB is not tempted to do so. And this improves Alice's overall payoff, by increasing her pre-time-τ_{n+1} payoff while leaving her post-τ_{n+1} payoff the same.

On the other hand, suppose $V_{\tau_n} < 0$: Shift everything prior to time t down by a common multiplicative factor $\lambda < 1$, but nofurther than $\lambda = \psi_{n+1}(\tau_n+1)/\psi_n(\tau_n+1)$. (Graphically, you lower the up-to-time τ_n ρ_t's down until the ψ_n curve just hits the ψ_{n+1} curve at $\tau_n + 1$. This has no impact on the triggering decision of any type of Evil Bob (the usual arguments apply), while it improves Alice's overall payoff by lowering her pre-τ_n losses. Note that when the constraint $\lambda \geq \psi_{n+1}(\tau_n+1)/\psi_n(\tau_n+1)$ is reached, the value of τ_n increases by 1, since at this point, the last time the constraint for B_n-EB binds is one time period later.

And in the knife-edge case that $V_{\tau_n} = 0$, either of the two suggested variations has no impact on Alice's overall payoff, so do one or the other; this is why the proposition gives the form of *an* optimal solution for Alice rather than *the* optimal solution.

Hence, $\psi_n(\tau_n) = \psi_{n+1}(\tau_n)$ can be assumed for all n, if Alice is optimizing.

The final step in the proof is to show that there is a time τ_N at which ρ_{τ_N} "reaches 1 precisely." That is, for some $t \geq t_N^*$, $\rho^{t_N^*}\gamma^{t_N^*-t} = 1$. We know that post time t_N^*, $\rho_t = \min\{1, \rho^{t_N^*}(\gamma_N)^{t_N^*-t}\}$. Let τ_N be the first time t that $\rho_{t+1}/\gamma_N > 1$. That is, $\rho_{\tau_N} \in (\gamma_n, 1]$. The unhappy possibility is that $\rho_{\tau_N} < 1$. But this is incompatible with Alice having optimized: We know that Alice's overall expected payoff at the optimum must be greater than zero, and if $\rho_{\tau_N} < 1$, she can proportionately increase her expected payoff by replacing each ρ_t for $t \leq \tau_N$ with ρ_t/ρ_{τ_N}. No Bob changes what he is doing, and her overall expected payoff increases proportionally. If she is optimizing, it must be that $\rho_{\tau_N} = 1$.

If you put all the pieces together, this proves the proposition. ∎

6. Only one "type" of Evil Bob[4]

If $N = 1$, Alice is looking for a single time, denoted τ_1 in the general notation system and abbreviated τ in this section. In this case, we can provide Alice with considerable assistance, as well as give fairly complete results. First, we adapt Proposition 3 to this special case.

Corollary 1. *If F_H is degenerate at a single value B and $\pi_G > 0$, then Alice's optimal solution takes the form*

$$\rho_t = \begin{cases} \gamma^{\tau-t}, & \text{for } t < \tau, \text{ and} \\ 1, & \text{for } t \geq \tau \end{cases},$$

where τ is a parameter of the solution chosen by Alice and $\gamma = \delta B/(B+1)$. This makes Evil Bob indifferent among triggering at times $0, 1, \ldots, \tau$; in this solution, Evil Bob chooses (on Alice's behalf) to trigger at $t = 0$ if $A > B$ and at $t = \tau$ if $A < B$. (If $A = B$, Alice is indifferent as to when Evil Bob triggers.)

Let $\beta = B/(B+1)$, so $\gamma = \delta\beta$. By simple algebra, $\beta/(1-\beta) = B$.

Suppose $A > B$ and Alice announces $\{\rho_t\}$ given by $\rho_t = \gamma^{\tau-t}$ for $t \leq \tau$ and $\rho_t = 1$ for $t > \tau$. Suppose Evil Bob triggers at time 0 (which he is happy

[4] I reiterate that, while the models differ in several respects, many of the results in this section are similar to results in Watson (1999a).

to do), while type-G Bob always treats Alice well. Then Alice's expected payoff

$$\pi_G\left[\sum_{s=0}^{\tau-1}\delta^s\gamma^{\tau-s}+\delta^\tau\frac{1}{1-\delta}\right]-\pi_H\gamma^\tau A=\pi_G\left[\sum_{s=0}^{\tau-1}\delta^s\delta^{\tau-s}\beta^{\tau-s}+\delta^\tau\frac{1}{1-\delta}\right]-\pi_H\gamma^\tau A$$

$$\pi_G\delta^\tau\left[\sum_{s=0}^{\tau-1}\beta^{\tau-s}+\frac{1}{1-\delta}\right]-\pi_H\gamma^\tau A=\pi_G\delta^\tau\left[\beta\frac{1-\beta^\tau}{1-\beta}+\frac{1}{1-\delta}\right]-\pi_H\gamma^\tau A.$$

Fixing the parameters in this expression, let $\mathcal{R}(\tau)$ denote this expected payoff (where $\mathcal{R}$ is mnemonic for, Bob triggers *right away*); rewriting again,

$$\mathcal{R}(\tau):=\pi_G\delta^\tau\left[B(1-\beta^\tau)+\frac{1}{1-\delta}\right]-(1-\pi_G)\gamma^\tau A. \tag{2.1}$$

where $\mathcal{R}$ is a mnemonic for, Bob triggers *right away*.)

And if $B>A$ and Alice chooses τ and $\{\rho_t\}$ as above. If Evil Bob triggers at time τ, which is in Alice's best interests, Alice's expected payoff is

$$\begin{aligned}\mathcal{L}(\tau):=&\sum_{s=0}^{\tau-1}\delta^s\gamma^{\tau-s}+\delta^\tau\left[\pi_G\frac{1}{1-\delta}-\pi_H A\right]=\delta^\tau\left[\beta\frac{1-\beta^\tau}{1-\beta}+\pi_G\frac{1}{1-\delta}-\pi_H A\right]\\&=\delta^\tau\left[B(1-\beta^\tau)+\frac{\pi_G}{1-\delta}-(1-\pi_G)A\right].\end{aligned} \tag{2.2}$$

where $\mathcal{L}$ is a mnemonic for "later." Of course, $\mathcal{R}\equiv\mathcal{L}$ if $A=B$.

Recall that, back on page 9, we said " The idea, at least at first blush, is to get Evil Bob to reveal himself at a lower and therefore less costly scale. (A second way that starting small can help Alice will be developed.)" If $A>B$, we have the first-blush idea: Alice wants Evil Bob to trigger right away and, indeed, she has scaled her engagements to get him to do so at a small initial scale. Of course, she pays a price for this: the smaller the scale on which she has Evil Bob reveal himself, the longer she must wait to get to full engagement with Bob, if Bob is indeed type G.

But if $A<B$, we have the "second way." By starting small, Alice essentially entices Evil Bob to wait to trigger. Until he does, he is treating her well, which is good for her. And when he triggers, it is at full scale, but in

the future so that discounting reduces the impact of being treated poorly on her overall payoffs, the impact of discounting also being good for her.

Comparative statics

Let $\mathcal{V}(\delta, A, B, \pi_G)$ be Alice's optimal expected payoff for the parameters δ, A, B, and π_G, and let $\tau^*(\delta, A, B, \pi_G)$ be her optimal time τ. We have:

Proposition 4. *If $\pi_G \in (0,1)$, then $\mathcal{V}$ is strictly increasing in δ and π_G. It is strictly decreasing in A. It is strictly decreasing in B as long as $A > B$, and it is strictly increasing in B as long $B > A$.*

Proof. Of course, $\mathcal{V} = \max\{\mathcal{R}(\tau); \tau = 0, 1, \ldots\}$ if $A > B$, and $\mathcal{V} = \max\{\mathcal{L}(\tau); \tau = 0, 1, \ldots\}$ if $B > A$, where the dependence on the four parameters is suppressed.

By immediate inspection, for each τ, both $\mathcal{R}(\tau)$ and $\mathcal{L}(\tau)$ are strictly increasing in π_G and strictly decreasing in A.

To show that $\mathcal{V}$ is strictly increasing in π_G, suppose (first) that $A > B$. Fix parameters δ, A, and B, and vary the parameter π_G, considering two values, $\check{\pi}_G > \hat{\pi}_G$. Let $\hat{\tau}^*$ be the optimal τ for a given parameters with $\hat{\pi}_G$, and let $\check{\tau}^*$ be the optimal τ for the parameters and $\check{\pi}_G$. We have

$$\mathcal{V}(\check{\pi}_G) = \mathcal{R}(\check{\tau}^*; \check{\pi}_G) \geq \mathcal{R}(\hat{\tau}^*; \check{\pi}_G) > \mathcal{R}(\hat{\tau}^*; \hat{\pi}_G) = \mathcal{V}(\hat{\pi}_G),$$

where $\mathcal{V}(\cdot)$ is $\mathcal{V}$ for the set of parameters with the value of π_G the argument, and similarly for the arguments of the $\mathcal{R}$ after the semi-colons. The first inequality in this sequence is because $\check{\tau}^*$ is optimal for $\check{\pi}_G$, and the second is because $\mathcal{R}$ is strictly increasing in π_G. Hence, if $A > B$, $\mathcal{V}$ is increasing in π_G. If $A < B$, the same argument will work, substituting $\mathcal{L}$ for $\mathcal{R}$ and enlisting the second paragraph of the proof.

To show that $\mathcal{V}$ is strictly decreasing in A, fix parameters except for A and suppose $\check{A} < \hat{A} < B$ (so $\mathcal{V}$ is in both cases given by $\mathcal{L}$. Letting $\check{\tau}^*$ and $\hat{\tau}^*$ be the optimal τ's for $\check{A}$ and $\hat{A}$, respectively,

$$\mathcal{V}(\hat{A}) = \mathcal{L}(\hat{\tau}^*; \hat{A}) < \mathcal{L}(\hat{\tau}^*; \check{A}) \leq \mathcal{L}(\check{\tau}^*; \check{A}) = \mathcal{V}(\check{A}).$$

The argument for $B \leq \check{A} < \hat{A}$ is similar, with $\mathcal{R}$ replacing $\mathcal{L}$. And for the case $\check{A} < B \leq \hat{A}$, proceed as in the last display, but throw into the sequence of inequaliities $\mathcal{L}(\hat{\tau}^*; \hat{A}) < \mathcal{R}(\hat{\tau}^*; \hat{A})$ and then finish with $\mathcal{R}$'s in place of the $\mathcal{L}$'s.

To show monotonicity of $\mathcal{V}$ in δ: First take the case $A \geq B$, and fix all parameters except for δ. Rewrite (2.1) as

$$\mathcal{R}(\tau) = \delta^\tau \left\{ \pi_B \left[B(1-\beta^\tau) + \frac{1}{1-\delta} \right] - (1-\pi_G)\beta^\tau A \right\}.$$

This is clearly increasing in δ for each τ, so if $\check{\delta} > \hat{\delta}$ and $\hat{\tau}$ is optimal for $\hat{\delta}$,

$$\mathcal{V}(\check{\delta}) \geq \mathcal{R}(\hat{\tau}; \check{\delta}) > \mathcal{R}(\hat{\tau}; \hat{\delta}) = \mathcal{V}(\hat{\delta}).$$

For the case $B > A$, we don't need to rewrite (2.2); it is immediate that $\mathcal{L}(\tau)$ is increasing in δ, and a similar argument works.

And for the behavior of $\mathcal{V}$ in B. For any set of parameters, with the dependence on B made explicit, write $\mathcal{R}^*(B) = \max_\tau \{\mathcal{R}(\tau; B)\}$ and $\mathcal{L}^*(B) = \max_\tau \{\mathcal{L}(\tau; B)\}$. Since for any set of parameters, $\mathcal{R}(\tau)$ is strictly positive for some (large enough) τ and asymptotes to 0 as $\tau \to \infty$, the definition of $\mathcal{R}^*(B)$ is fine for any B. But care is needed for $\mathcal{L}^*$; we are guaranteed that there is some best τ if $B > A$. In fact, as long as $\pi_G < 1$, it is easy to show that there is some $\epsilon > 0$ such that $\mathcal{L}(\tau) > 0$ for large enough τ, as long as $B > A - \epsilon$, so take the domain of $\mathcal{L}^*$ to be $B > A - \epsilon$ for that ϵ.

I assert that $\mathcal{R}^*(B)$ is strictly decreasing in B. Take $\check{B} < \hat{B}$, and let $\hat{\tau}$ be such that $\mathcal{R}^*(\hat{B}) = \mathcal{R}(\hat{\tau}; \hat{B})$. Suppose that Alice, faced with $\check{B}$, announces $\{\rho_t\}$ given by

$$\rho_t = \min \left\{ 1, \frac{(\gamma_{\hat{B}})^{\hat{\tau}}}{(\gamma_{\check{B}})^t} \right\}.$$

That is, she starts with the same ρ_0 as was optimal for $\hat{B}$, but rises more quickly. If $\check{B}$-EB triggers at time 0, the cost to Alice is the same as the cost of $\hat{B}$-EB triggering at time 0, and her payoffs from type-G Bobs is at every date larger than against the sequence of scales that gives $\mathcal{R}(\hat{\tau}; \hat{B})$, so she is

strictly better off. In particular, her expected payoff is strictly positive. Now by scaling all the scales up, so the sequence hits 1 precisely at some $\check{\tau}$, we have

$$\mathcal{R}^*(\check{B}) \geq \mathcal{R}(\check{\tau}; \check{B}) > \mathcal{R}(\hat{\tau}; \hat{B}) = \mathcal{R}^*(B).$$

And I assert that $\mathcal{L}^*(B)$ is strictly increasing in B, for B large enough so that $\mathcal{L}^*(B) > 0$. Again take $\check{B} < \hat{B}$, and let $\check{\tau}$ be such that $\mathcal{L}(\check{\tau}; \check{B}) = \mathcal{L}^*(\check{B})$. Suppose Alice declares $\{\rho_t\}$ given by $\rho_t = 1$ for $t \geq \check{\tau}$ and $\rho_t = (\gamma_{\hat{B}})^{\check{\tau}-t}$ for $t < \check{\tau}$. $\hat{B}$-EB will happily wait until $\check{\tau}$ to trigger, so the triggering cost to Alice will be the same as if she faced $\check{B}$-EB who triggered at the same time (both at scale 1). And, up to time $\check{\tau}$, she gets larger payoffs from all types. Hence

$$\mathcal{L}^*(\hat{B}) \geq \mathcal{L}(\check{\tau}; \hat{B}) > \mathcal{L}(\check{\tau}; \check{B}) = \mathcal{L}^*(\check{B}).$$

But since $\mathcal{V} = \mathcal{R}^*$ for $A > B$ and $= \mathcal{L}^*$ for $B > A$ (they are identical at $A = B$, which the reader can verify by inspection of 2.1 and 2.2), we have the result. ∎

The most interesting part of Proposition 4 is the U-shape of $\mathcal{V}$ in B. (In fact, it is more of a V-shape, as there is a distinct kink at $A = B$.) The proof given provides the intuition. For $A > B$, Alice is using the *impatience* of Evil Bob to get him to trigger at time 0. Hence, the more impatient he is—the smaller is his B—the more she can use that impatience against him. But for $B > A$, Alice is relying on Evil Bob's *patience* to wait until τ^* to trigger. Hence, the more patient he is—the bigger is his B—the better for her.

To obtain partial comparative statics results for τ^*, we proceed as follows. While $\mathcal{R}$ and $\mathcal{L}$ as defined in equations 2.1 and 2.2 have domain $\{0, 1, 2, \ldots\}$, we can use those formulas to expand the domain of definition of the functions to all of $[0, \infty)$. To distinguish between these cases, I'll use ξ to denote the continuous argument of the functions. We know (Proposition 1(d)) that if $A \geq B$, $\mathcal{R} > 0$ for some argument (for any set

of parameters, as long as $\pi_G > 0$), while if BA, $\mathcal{L} > 0$ for some argument. (In fact, $\mathcal{R} > 0$ for some argument even if $B > A$, but if $A > B$, $\mathcal{L}(\xi) < 0$ for all ξ is possible.) Inspection of 2.1 and 2.2 immediately shows that $\lim_{\xi\to\infty} \mathcal{R}(\xi) = \lim_{\xi\to\infty} \mathcal{L}(\xi) = 0$. And:

Lemma 2. *If $A \geq B$, the derivative of $\mathcal{R}$ in ξ, $\mathcal{R}'$, is either everywhere nonpositive or is strictly positive and then negative, passing through 0 exactly once (for any given set of parameters). If $B > A$, the derivative of $\mathcal{L}$ in ξ, $\mathcal{L}'$, is either everywhere nonpositive or is strictly positive and then negative, passing through 0 exactly once. Hence, in the case $A > B$, if $\mathcal{R}'(0) \leq 0$, $\tau = 0$ is optimal for Alice, while if $\mathcal{R}'(0) > 0$ and ξ^* is the unique solution to $\mathcal{R}'(\xi^*) = 0$, the optimal τ for Alice is either $\lfloor \xi^* \rfloor$ or $\lfloor \xi^* \rfloor + 1$. ($\lfloor x \rfloor$ denotes the integer part of x.) And, in the case $B > A$, the same is true for Alice's optimal τ, with $\mathcal{L}$ replacing $\mathcal{R}$.*

Proof. Suppose A > B. Rewrite $\mathcal{R}(\xi)$ as

$$\pi_G \delta^\xi \left[B + \frac{1}{1-\delta} \right] + \gamma^\xi \left[\pi_G A - A + \pi_G B \right],$$

and we have

$$\begin{aligned} \mathcal{R}'(\xi) &= \ln(\delta)\pi_G \delta^\xi \left[B + \frac{1}{1-\delta} \right] + \ln(\gamma)\gamma^\xi \left[\pi_G A - A + \pi_G B \right] \\ &= \delta^\xi \left\{ \ln(\delta)\pi_G \left[B + \frac{1}{1-\delta} \right] + \ln(\gamma)\beta^\xi \left[\pi_G A - A + \pi_G B \right] \right\}. \end{aligned}$$

Hence

$$\frac{\mathcal{R}'(\xi)}{\delta^\xi} = \ln(\delta)\pi_G \left[B + \frac{1}{1-\delta} \right] + \ln(\gamma)\beta^\xi \left[\pi_G A - A + \pi_G B \right].$$

Since both $\ln(\delta)$ and $\ln(\gamma)$ are negative, this is the sum of a strictly negative term that is constant in ξ, plus a term of undetermined sign that monotonically decreasing as ξ increases, asymptoting to zero. If $\mathcal{R}'(0) > 0$ (which is $\ln(\delta)[B + 1/(1-\delta)] + \ln(\gamma)[\pi_G A - A - \pi_G B] > 0$), $\mathcal{R}'(\xi)/\delta^\xi = 0$ at exactly one point; viz., the solution in ξ to

$$\beta^\xi = \frac{\ln(\delta)\left[B + 1/(1-\delta)\right]}{\ln(\gamma)\left[\pi_G A - A - \pi_G B\right]}. \tag{3.1}$$

And since the sign of $\mathcal{R}'(\xi)/\delta^\xi$ is the same as the sign of $\mathcal{R}'(\xi)$, the same is true for $\mathcal{R}'$. That is, either $\mathcal{R}'$ is everywhere negative or it starts out strictly positive and crosses 0 in exactly one point, to become and remain strictly negative.

And for the case $B > A$, we have

$$\begin{aligned}
\mathcal{L}(\xi) &= \delta^\xi \left[B + \frac{\pi_G}{1-\delta} - A + \pi_G A \right] - B\gamma^\xi, \quad \text{hence} \\
\mathcal{L}'(\xi) &= \ln(\delta)\delta^\xi \left[B + \frac{\pi_G}{1-\delta} - A + \pi_G A \right] - \ln(\gamma) B\gamma^\xi \\
&= \delta^\xi \left\{ \ln(\delta) \left[B + \frac{\pi_G}{1-\delta} - A + \pi_G A \right] - \ln(\gamma) B\beta^\xi \right\}.
\end{aligned}$$

$\mathcal{L}'(\xi)/\delta^\xi$ is the sum of strictly negative term ($\ln(\delta) < 0$ and, since $B > A$, the term in the square brackets is strictly positive) less a strictly negative term (that is, plus a positive term) whose magnitude is monotonically decreasing in ξ. Hence $\mathcal{L}'(\xi)$ is either everywhere negative (if the term inside the curly brackets is negative at $\xi = 0$ or starts positive, passes through zero at one point, and thereafter is strictly negative. And, if it is ever zero, it is zero where

$$\beta^\xi = \frac{\ln(\delta)\left[B + \pi_G/(1-\delta) - A + \pi_G A\right]}{\ln(\gamma)B}. \tag{3.2}$$

The rest of the lemma is straightforward. ∎

Proposition 5. *$\tau^*(\delta, A, B, \pi_G)$ is weakly decreasing π_G and weakly increasing in A.*

Proof. By inspection, if $A > B$, $\mathcal{R}'(\xi; \delta, A, B, \pi_G)$ is increasing in A for all the other arguments fixed, and it is decreasing in π_G for all the other arguments fixed. And, if $B > A$, the same is true of $\mathcal{L}'$. Hence, the solutions to $\mathcal{R}'(\xi) = 0$ when $A > B$ ($\xi = 0$ if $\mathcal{R}'(0) < 0$) and $\mathcal{L}'(\xi) = 0$ when $B > A$ ($\xi = 0$ if $\mathcal{L}'(0) < 0$) are increasing in A and decreasing in π_G. Moreover, suppose $\hat{A} > \check{A} > B$. Let $\hat{\xi}$ be the solution to $\mathcal{R}'(\xi) = 0$ for parameter $\hat{A}$,

and similarly for $\check{\xi}$. What happens if $\lfloor\hat{\xi}\rfloor = \lfloor\check{\xi}\rfloor$? If $\check{\tau}^*$ is $\lfloor\check{\xi}\rfloor + 1$, it would be because

$$\int_{\lfloor\check{\xi}\rfloor}^{\lfloor\check{\xi}\rfloor+1} \mathcal{R}'(\xi; \check{A}) > 0;$$

that is, $\mathcal{R}$ rises more between $\lfloor\check{\xi}\rfloor$ and $\check{\xi}$ then it falls between $\check{\xi}$ and $\lfloor\check{\xi}\rfloor + 1$ (all this for fixed parameters and $A = \check{A}$). But then, if we substitute $A = \hat{A}$, since $\mathcal{R}'$ rises for every ξ, the same integral inequality holds, hence $\hat{\tau}^* = \lfloor\check{\xi}\rfloor$. This shows that τ^* is weakly rising in A, for $A > B$. The same argument works for $B > A$, and you can "bridge" across the two cases by interposing the case $A = B$ and employing transitivity.

And the same argument works for how τ^* varies with π_A (except that the "bridging" step is unnecessary). ∎

It is my strong belief (based on numerical examples) that τ^* is (weakly) increasing in B and in δ, but I am unable to produce proofs that these conjectures are true at this time.

Alice "solves" the one-B case numerically

We know, and Alice knows, that the τ^* for a given problem is either $\lfloor\xi^*\rfloor$ or $\lfloor\xi^*\rfloor + 1$, where ξ^* is the solution to $\mathcal{R}'(\xi) = 0$ or $\mathcal{L}'(\xi) = 0$, as appropriate given A and B, (and with the proviso that if $\mathcal{R}(0)$ or $\mathcal{L}(0)$ is less than zero, $\tau^* = 0$). Hence, from 3.1 and 3.2, she can quickly solve for ξ^* and then try the adjacent integer values. For the case $A > B$, she solves

$$\xi^* = \ln\left(\frac{\ln(\delta)\big[B + 1/(1-\delta)\big]}{\ln(\gamma)\big[\pi_G A - A - \pi_G B\big]}\right) \Big/ \ln(\beta),$$

and for $B > A$,

$$\xi^* = \ln\left(\frac{\ln(\delta)\big[B + \pi_G/(1-\delta) - A + \pi_G A\big]}{\ln(\gamma)B}\right) \Big/ \ln(\beta).$$

Example #	1	2	3	4	5	6
δ	0.9	0.9	0.9	0.9	0.9	0.9
A	8	8	8	9.5	12	15
B	5	5	5	5	5	5
π-sub-G	0.35	0.5	0.95	0.35	0.5	0.9
ξ	7.04789	4.72444	-0.07291	7.76794	6.19582	1.06152
int[ξ]	**7**	4	**0**	7	**6**	**1**
Value of $\mathcal{R}$	**1.583**	2.864	**9.100**	1.453	**2.473**	**8.108**
int[ξ] + 1	8	**5**	N.A.	**8**	7	2
Value of $\mathcal{R}$	1.564	**2.886**		**1.467**	2.453	7.967

(a) Examples with $A > B$

Example #	7	8	9	10	11	12
δ	0.9	0.9	0.9	0.9	0.9	0.9
A	5	5	5	6	7	9
B	10	10	10	10	10	10
π-sub-G	0.25	0.55	0.95	0.32	0.5	0.89
ξ	8.16081	3.80720	-0.11173	7.72627	5.29340	0.64529
int[ξ]	**8**	3	**0**	7	**5**	0
Value of $\mathcal{L}$	**1.758**	4.182	**9.250**	1.908	**3.124**	7.910
int[ξ] + 1	9	**4**	N.A.	**8**	6	**1**
Value of $\mathcal{L}$	1.747	**4.212**		**1.918**	3.112	**7.937**

(b) Examples with $B > A$

Table 1. A few examples of one-B Evil Bob

(In either case, if the resulting ξ^* is negative, this means $\tau^* = 0$. Table 1 provides some examples, with the optimal τ^* and value of $\mathcal{V}$ highlighted by bold-face and slightly larger font.

How do the payoffs of Evil Bob and type-G Bobs change with B

The effect of B on $\mathcal{V}$, as already noted, is the most interesting: $B \to \mathcal{V}$ is U-shaped, hitting a minimum at $B = A$, increasing as B recedes from A on either side. And while it isn't a proposition, yet, my strong conjecture is that $B \to \tau^*$ is weakly increasing.

The impact that B has on the payoff of Evil Bob and on type-G Bobs is quite complex. The mere fact that B increases is surely good news for Evil Bob, since B is his reward for triggering. But it isn't quite that simple, and it certainly isn't simple how B impacts the payoff of type-G Bobs. If the

conjecture is right that increasing B weakly increases τ^*, that is bad news for both Evil Bob and the type-G Bobs, as (i) it weakly increases the number of "discounts" that Alice applies, and (ii) it delays for type-G Bobs the time at which they (finally) reach full scale. But, on the other hand, increasing B decreases γ, which is good news for both type-G Bobs and Evil Bob. Assuming the conjecture that τ^* is weakly increasing in B is correct, then, one expects that increasing B will give a "sawtooth" for type-G Bobs, with their payoff rising in B as long as τ^* stays fixed, but jumping down at points where τ^* increases by 1. Evil Bob's payoff should also follow this pattern, although in numerical examples it may be harder to see (unless we increase B on a very fine scale), since there is the additional effect that bigger B is, directly, good news for Evil Bob.

All this can be seen numerically. Table 2 provides data for $\delta = 0.9$, $A = 7$, $\pi_G = 0.5$, and $B = 2, 3, \ldots, 13$. Note in particular that the payoff to Evil Bob declines from $B = 11$ to $B = 12$ because of the step up in τ^*, which has greater effect over this interval than the increase in γ and B itself.

B	2	3	4	5	6	7	8	9	10	11	12	13
ξ	3.18389	3.66434	4.01535	4.28542	4.50061	4.67656	4.91223	5.11574	5.2934	5.44997	5.58904	5.71345
τ^*	3	4	4	4	5	5	5	5	5	5	6	6
$\mathcal{V}$ --- Alice's ExNPV	3.4020	3.2267	3.1146	3.0223	2.9481	2.8991	2.9882	3.0620	3.1242	3.1771	3.2293	3.2771
Payoff to EB	0.4320	0.6228	1.0750	1.5820	1.6392	2.1201	2.6214	3.1381	3.6665	4.2040	3.9452	4.4288
Payoff to type-G	8.3160	7.9065	8.1104	8.2595	7.8086	7.9183	8.0074	8.0812	8.1433	8.1963	7.7466	7.7943

Table 2. Varying B. (See the text for explanation.)

7. Two types of Evil Bob

While the case of one type of Evil Bob is "solvable," albeit numerically, as soon as Evil Bob comes in two flavors, Alice is stuck with (more or less) brute force searches to find her best pair $\{\tau_1^*, \tau_2^*\}$ (which then determine the sequence $\{\rho_t\}$ via Proposition 3).[5] But, at the same time—and perhaps exhibiting why there are no easy answers—the case of two flavors of Evil Bob provides some provocative phenomena. So, in this section, I provide some examples.

[5] Although see the discussion of "single-peakedness" following.

The structure of Alice's problem is in most respects a standard screening problem, and in such problems, solutions are often constructed from "the bottom up." At least, in a separating equilibrium, one first solves without constraint for the worst type, then for the next worst type constrained to separate itself from the worst type, and so forth. But that technique works when one thinks of the uninformed side of the market as competitive.

Here, however, we are thinking of the uninformed party, Alice, as a Stackelberg leader. How she should handle the worst type of Evil Bob from her perspective is entirely bound up in how she plans to handle other types of Evil Bob.[6] Indeed, it is not clear from her perspective whether worse types of Bob are those with bigger Bs or smaller Bs or, perhaps, those with Bs that are closest to A (on either side): One might think that Evil Bob with a bigger B has more incentive to hurt her, so this is a worse type. But, as showed in Proposition 4, Alice prefers Evil Bob's whose B is distant from her parameter A.

Label the two values of B for Evil Bob B_1 and B_2, with $B_2 > B_1$. Let $\gamma_n = \delta B_n/(B_n + 1)$. Alice is looking for the best two times τ_1 and τ_2, and given these times, she will set

$$\rho_t = \begin{cases} 1, & \text{if } t \geq \tau_2, \\ (\gamma_2)^{\tau_2 - t}, & \text{for } t \text{ between } \tau_1 \text{ and } \tau_2, \text{ and} \\ (\gamma_2)^{\tau_2 - \tau_1} \cdot (\gamma_1)^{\tau_1 - t}, & \text{for } t = 0, 1, \ldots, \tau_1. \end{cases}$$

If $A \geq B_2 > B_1$, then B_2-EB triggers at time τ_1 and B_1-EB triggers at time $\tau_0 = 0$. If $B_2 > A \geq B_2$, then B_2-EB triggers at time τ_2 while B_1-EB still triggers at time 0. And if $B_2 > B_1 > A$, then B_2-EB triggers at time τ_2 and B_1-EB triggers at time τ_1. With these formulae, and with the parameters π_G and ϕ_1 specified (ϕ_1 being the conditional probability that EB is B_1-EB, given that he is in fact evil), Alice's expected payoff as a function of τ_1 and τ_2 is easily (albeit tediously) computed, and we can present a table of values (given all the parameter values) over which she optimizes. In fact, it is convenient to have the rows and the columns in the table be τ_1 and $\xi_2 = \tau_2 - \tau_1$, and that is what we do. See Table 3 for a "typical" case.

[6] This is hardly new. In classic screening theory with a monopolist doing the screening, this issue arises.

τ_1 \ ξ_2	0	1	2	3	4	5
0	1.5000	1.9973	2.3577	2.6067	2.7656	2.8520
1	2.3400	2.6543	2.8634	2.9877	3.0442	3.0473
2	2.7743	2.9672	3.0775	**3.1220**	3.1146	3.0669
3	2.9479	3.0608	3.1075	3.1021	3.0561	2.9792
4	2.9576	3.0182	3.0248	2.9892	2.9213	2.8290
5	2.8673	2.8942	2.8762	2.8235	2.7444	2.6459

Table 3. An example with two flavors of Evil Bob. This table gives Alice's expected payoffs as a function of τ_1 and $\xi_2 = \tau_2 - \tau_1$, for the case $\delta = 0.9$, $A = 7$, $B_1 = 3$, $B_2 = 25$, $\pi_G = 0.5$, and $\phi_1 = 0.7$. The optimal combination, indicated by the boldface and slightly larger font, is $\tau_1 = 2$ and $\xi_2 = 3$ (hence $\tau_2 = 5$). Since $B_1 < A$, B_1-EB triggers at time 0; since $B_2 > A$, B_2-EB triggers at time $\tau_2 = 5$, at which point ρ_t has reached 1.

This example is for the case $\delta = 0.9$, $A = 7$, $B_1 = 3$, $B_2 = 25$, $\pi_G = 0.5$, and $\phi_1 = 0.7$. For each combination of τ_1 between 0 and 5 and ξ_2 between 0 and 5, Alice's expected payoff is provided. Note that, among all these possibilities, the best for her is $\tau_1 = 2$ and $\xi_2 = 3$, hence $\tau_2 = 5$, giving her a payoff of 3.122.

Compare this with how Alice does if she faced one of these two Evil Bobs alone. If Evil Bob has $B = 3$ for sure (with $\pi_G = 0.5$), Alice would set $\tau = 4$ and have an expected payoff of 3.277. If Evil Bob has $B = 25$ for sure (with $\pi_G = 0.5$), she would set $\tau = 7$ and accrue an expected payoff of 3.588. She does worse facing the two types than facing either type alone, since her best scheme fits neither one perfectly.

Table 3 shows a limited range of values for τ_1 and ξ_2. So how can we be sure that there isn't a better arrangement for Alice outside the bounds of the table?

1. In fact, in my numerical analysis, I looked over the ranges $0 \leq \tau_1 \leq 13$ and $0 \leq \xi_2 \leq 13$, and there was nothing better in that expanded range.

2. If Alice wanted to be sure, she could run out τ_1 so far that, by a combination of reduced scale and discounting, it is impossible that she does better than 3.122.

But, in addition, note that looking across each row and each column, the expected payoffs for Alice are single peaked, reminiscent of what happens

in the case of a single value of B. In the general (finitely many B) case, simple computations show the following: Reframe Alice's problem as one of choosing $\xi_1 = \tau_1 - \tau_0, \xi_2 = \tau_2 - \tau_1, \ldots, \xi_N = \tau_N - \tau_{N+1}$. Let $\mathcal{V}(\xi_1, \xi_2, ..., \xi_N)$ be Alice's expected payoff as a function of her selection of the ξ_n. Fix ξ_n for all n except n' and look at how $\mathcal{V}$ varies in $\xi_{n'}$ alone, treating $\xi_{n'}$ as a continuous rather than discrete variable. Then is it relatively easy to show that $\mathcal{V}(\xi_{n'})$ (holding the other ξ_n and the various parameters fixed) takes the form

$$K_1 \cdot (\gamma_{n'})^{\xi_{n'}} + K_2 \cdot \delta^{\xi_{n'}},$$

for constants K_1 and K_2 (that depend on the other ξ_n). Roughly speaking, K_1 reflects Alice's expected payoff from what happens from time 0 up to time $\tau_{n'-1}$ if $\xi_{n'} = 0$, while K_2 reflects Alice's expected payoff from what happens from time τ_n out to $t = \infty$ if $\xi_n = 0$. This isn't precisely true because, while $\mathcal{V}$ has the form indicated, we must account for what happens between times $\tau_{n'-1}$ and τ_n: If $A > B_{n'}$, then $B_{n'}$ triggering at τ_{n-1} clearly belongs to K_1, but the impact of $\xi_{n'}$ on what Alice gets from the type G Bobs and the B_n-EBs for $n > n'$ is "split" between K_1 and K_2. If $B_{n'} > A$, things split up differently.

But, except for this "fudge," it is quite intuitive why K_1 is multiplied by $(\gamma_{n'})^{\xi_{n'}}$ and K_2 is multiplied by $\delta^{\xi_{n'}}$: $\xi_{n'}$ doesn't affect the timing of anything that happens up to time $\tau_{n'-1}$, but it does reduce the scale on which those things happen by $(\gamma_{n'})^{\xi_{n'}}$. And while $\xi_{n'}$ doesn't affect the scale of what happens after time $\tau_{n'}$, but it does delay those things by $\xi_{n'}$, so they are discounted by an additional $\delta^{\xi_{n'}}$.

And *if* the $\{\xi_n\}$ that Alice is looking at are such that $K_2 > 0$—which makes sense, since she is presumably looking in regions where her overall expected payoff is positive and, post $\tau_{n'}$, she has shed some of the Evil Bobs that are bad for her—then the proof that $\mathcal{V}$ (viewed as a function of the continuous variable $\xi_{n'}$) is single-peaked that was employed in the case of a single B works just as well here.

This fact isn't quite as iron-clad useful as in the case of a single B. With a single B, where one can only increase or decrease a single variable, the

optimal integer-constrained value is either immediately to the left or right of where the derivative passes through zero (from above to below). With multiple dimensions, one doesn't quite have a guarantee that, if one knew where all the partial derivatives of $\mathcal{V}$ are zero, the integer-constrained optimum would be at an "adjacent" integer-lattice value. And while, for the single B case, formulae for the where the single derivative of $\mathcal{V}$ is zero can be written down (which, of course, we did last section), with multiple B: (i) writing down the partial derivatives is quite painful, and (ii) solving for where they are simultaneously equal to zero is even more painful.

That said, it is certainly the case that, in doing a numerical search as in Table 3, Alice can be comforted that is she sees integer values of τ_1 and ξ_2 such that her expected payoff falls off in all directions, she can probably stop searching; she probably has found the best combination for her.

In Table 4, three numerical examples are presented. In all three, $\delta = 0.9$ and $\pi_G = 0.5$. Panel a provides a typical example in which $A > B_2 > B_1$; specifically, $A = 10$, $B_1 = 2$, and $B_2 = 7$. Each row give results for a different value of ϕ_1, ranging from $\phi_1 = 0$ to $\phi_1 = 1$. For each value of ϕ_1 (and the other parameters), τ_1^* and ξ_2^* (the optimal values for τ and ξ) are given, followed by: $\mathcal{V}$, Alice's expected payoff; $\mathcal{W}_1$, the payoff to B_1-EB; $\mathcal{W}_2$, the payoff to B_2-EB; and $\mathcal{W}_G$, the payoff to any type-G Bob. Panels b and c provide the same data for $B_1 = 3 < A = 7 < B_2 = 25$ (the example of Table 3) and $A = 5 < B_1 = 7 < B_2 = 30$, respectively

There are several things to note:

- For the entries for $\phi_1 = 0$ and 1 give the optimal values if only B_2-EB or, respectively, B_1-EB were present.
- The payoffs to Bob (of any type) only depend on τ_1^* and ξ_2^*. And these are for-sure payoffs, given Alice's strategy.
- In all three cases, for $\phi_1 = 0.1$, Alice chooses τ_1^* and ξ_2^* as if B_1-EB was not present. In panel c, this extends to $\phi_1 = 0.2$. And in panels a and c, symmetric effects are seen for ϕ_1 close to 1. (In panel b, Alice "ignores" B_2-EB somewhere between $\phi_1 = 0.96$ and 0.97.
- In panel a, where $B_1 < B_2 < A$, B_1-EB always triggers at $t = 0$, while B_2-EB triggers at time τ_1^*. Hence we have "full pooling" for $\phi_1 = 0.1$

ϕ_1	τ_1^*	ξ_2^*	$\mathcal{V}$	$\mathcal{W}_1$	$\mathcal{W}_2$	$\mathcal{W}_G$
0	0	6	2.490	0.477	1.670	7.365
0.1	0	6	2.490	0.477	1.670	7.365
0.2	0	6	2.490	0.477	1.670	7.365
0.3	1	5	2.491	0.363	1.726	7.308
0.4	1	5	2.527	0.363	1.726	7.308
0.5	1	5	2.564	0.363	1.726	7.308
0.6	2	4	2.631	0.277	1.835	7.200
0.7	2	3	2.712	0.352	2.330	7.709
0.8	3	2	2.831	0.268	2.528	7.510
0.9	3	1	2.959	0.340	3.211	7.943
1	4	0	3.159	0.259	3.540	7.614

a. $A = 10$, $B = 2$, $B = 7$

ϕ_1	τ_1^*	ξ_2^*	$\mathcal{V}$	$\mathcal{W}_1$	$\mathcal{W}_2$	$\mathcal{W}_G$
0	0	7	3.588	1.090	9.087	7.654
0.1	0	7	3.485	1.090	9.087	7.654
0.2	0	6	3.382	1.260	10.500	8.100
0.3	1	5	3.301	0.983	10.593	8.008
0.4	1	5	3.237	0.983	10.593	8.008
0.5	1	5	3.174	0.983	10.593	8.008
0.6	2	4	3.144	0.767	10.761	7.840
0.7	2	3	3.122	0.886	12.435	8.233
0.8	3	2	3.131	0.691	12.702	7.965
0.9	3	1	3.161	0.798	14.677	8.286
1	4	0	3.227	0.623	15.057	7.907

b. $A = 7$, $B_1 = 3$, $B_2 = 25$

ϕ_1	τ_1^*	ξ_2^*	$\mathcal{V}$	$\mathcal{W}_1$	$\mathcal{W}_2$	$\mathcal{W}_G$
0	0	6	4.176	3.056	13.096	8.162
0.1	0	6	4.057	3.056	13.096	8.162
0.2	0	6	3.939	3.056	13.096	8.162
0.3	1	4	3.835	3.172	15.084	8.536
0.4	1	4	3.744	3.172	15.084	8.536
0.5	2	3	3.668	2.868	15.177	8.443
0.6	2	3	3.599	2.868	15.177	8.443
0.7	3	1	3.557	2.978	17.581	8.663
0.8	4	0	3.541	2.692	17.782	8.462
0.9	4	0	3.541	2.692	17.782	8.462
1	4	0	3.541	2.692	17.782	8.462

c. $A = 5$, $B_1 = 7$, $B_2 = 30$

Table 4. Three examples. For three examples, Alice's optimal values of τ_1 and ξ_2 as well as her expected payoff $\mathcal{V}$ and the payoffs of B_1-EB, B_2-EB, and any type-G Bob ($\mathcal{W}_1, \mathcal{W}_2,$ and $\mathcal{W}_G$, respectively, are provided for a range of values of ϕ_1. In all cases, $\delta = 0.9$ and $\pi_G = 0.5$. The values of A, B_1, and B_2 are provided in the panel legends.

and 0.2. In panel c, where $A < B_1 < B_2$, B_1-EB always triggers at $t = \tau_1^*$ and B_2-EB triggers at $\tau_1^* + \xi_2^*$. Hence "full pooling" occurs when $\xi_2^* = 0$, which happens somewhere between $\phi_1 = 0.7$ and 0.8.

- When $B_1 < A < B_2$, as in panel b, B_1-EB triggers at $t = 0$ and B_2-EB triggers at $\tau_1^* + \xi_1^*$, so full pooling requires that $\tau_1^* + \xi_1^* = 0$. For the parameter values in panel b, this doesn't happen. And even if, say, $\phi_1 = 0.1$, so Alice behaves as if B_2-EB did not exist, the two types of Evil Bobs do not "pool," insofar as they trigger at different times. Similar remarks apply to panel a for $\phi_1 = 0.9$ and panel c for $\phi_1 = 0.1$ and 0.2.
- In panel b for ϕ_1 from 0.3 to 0.9, while B_1-EB triggers at time 0 and B_2-EB triggers at time $\tau_1^* + \xi_1^*$, the sequence $\{\rho_t\}$ that is optimal for Alice "breaks" at an intermediate time. Roughly put, as we move back in time from $\tau_1^* + \xi_1^*$, where ρ_t first reaches one, the ρ_t's decline slowly, presumably to increase Alice's take from type-G Bobs and B_2-EB. But at time τ_1^*, the decline is steeper, presumably to decrease the cost to Alice of B_1-EB's triggering at time 0.
- Comparing in each case the sum $\tau_1^* + \xi_1^*$ for the extreme values of ϕ_1, 0 and 1, it is always the case that the larger B_2 gives a higher value than the smaller B_1. While I was unable to prove that τ^* is weakly increasing in B for the single-B case, in all numerical examples I've looked at, this is true, and I hypothesize that the result is general.
- Between the two extremes $\phi_1 = 0$ and 1, Alice optimally moves (in discrete steps) from the $\tau_1^* = 0$ regime when $\phi_1 = 0$ to the $\xi_2^* = 0$ regime when $\phi_1 = 1$. In these three examples (and all examples I've computed), B_2-EB's payoff is monotonically increasing in ϕ_1; the more Alice is concerned with B_1-EB and tailors her choice to him, the better it is for B_2-EB. This happens because Alice's concern for B_1-EB causes her to reduce ρ_t for $t < \tau_1^*$, which reduces the cost to B_2-EB as he waits for his time to trigger.
- But for both B_1-EB and type-G Bob, payoffs are *not* monotone in ϕ_1. There are two forces that compete: As ϕ_1 increases and Alice is increasingly tailoring $\{\rho_t\}$ to B_1-EB, τ_1^* increases. This both reduces the scale

at which B_1-EB triggers,[7] but it also sometimes causes a downward revision in the sum $\tau_1^* + \xi_2^*$, which is good for both B_1-EB and type-G Bobs (and is very good for B_2-EB).

8. Commitment by Alice?

We've assumed throughout that Alice can commit at the outset to the sequence of scales $\{\rho_t\}$. There are reasons why such a commitment might be credible; for instance, if Alice deals with many Bobs through time, with different starting dates, she might be maintaining a reputation for how she behaves in each relationship.

And, as long as the probability of Good Bob is large enough relative to the probability that Bob is saintly, one can construct a sequential equilibrium in which it is always in Alice's interests to carry out her announced sequence of scales: Imagine that she announces the sequence of scales and then, if she ever deviates, Good Bob (and Evil Bob, to the extent that there is positive probability that the possibility that Bob is evil remains positive) "infers" that Alice plans never to engage in the future, hence his best response is to trigger while he can. Saintly Bob would continue to treat Alice well, even anticipating that Alice will never engage again, but as long as ψ, the probability that Bob is good, conditional on him being type-G, is such that $(1-\psi)/(1-\delta) < A\psi$, Alice, threatened by these conveniently specified out-of-equilibrium beliefs by Good Bob, has no interest in deviating from her original announcement. (And, of course, she is happy that he has these beliefs.)

But this technical fix is not entirely palatable, given that Alice will often reach a position in which, ex post, her incentives to deviate from her preannounced $\{\rho_t\}$ are strong.

- Suppose there is only one flavor of Evil Bob, with a parameter $B < A$. Suppose that Alice's best (with commitment) announcement corresponds to $\tau^* \geq 2$. Evil Bob, according to the equilibrium, triggers at

[7] For cases where $B_1 > A$, this isn't so. Instead, it delays the time at which B_1-EB triggers. But remember that B_1-EB is indifferent between triggering at 0 and τ_1^*, so we can evaluate his payoff by seeing what he would get if he triggered at time 0, in which case the argument given is valid.

time 0, so if Bob does not trigger at time 0, Alice infers that Bob must be type G. But if he is type G, Alice (and Bob) would prefer to set $\rho_1 = 1$. And, of course, if Evil Bob anticipates that this will happen, and $\rho_0 = \gamma^2$, he will not trigger at time 0.

If $B < A$ and $\tau^* = 1$, this issue doesn't arise, as Alice's plan is to set $\rho_1 = 1$. This suggests how we might construct an equilibrium in which Alice faces no dilemma of this sort. For π_G sufficiently close to 1, Alice ignores the possibility of Evil Bob and sets $\rho_0 = 1$. Let π_G^0 be smallest value of π_G for which this is true. And let π_G^1 be the smallest value of π_G for which $\tau^* = 1$ is optimal for Alice. The problem arises if $\pi_G < \pi_G^1$. Suppose that this is so and that, in particular, for the given π_G, $\tau^* = 2$. Then imagine that Alice sets $\rho_0 = \gamma^2$, but instead of triggering with certainty at time 0, Evil Bob randomizes between triggering at time 0 (or at time 1), in a manner that, if he doesn't trigger, causes Alice to revise her probability that Bob is evil down to $1 - \pi_G^1$. Then, at time 1, Alice is happy to announce $\rho_1 = \gamma$, and everything proceeds nicely. Of course, Alice's payoffs are lowered because of this. But, at least, she isn't tempted to set $\rho_1 = 1$. This will work for at least some $\pi_G < \pi_G^1$; and it suggests how we might extend to multiple periods of randomization by Evil Bob, each period causing Alice's posterior probability assessment that Bob is evil to rise enough so she doesn't have an incentive to deviate.

- The other possibility is that $B > A$. Suppose this is so and $\tau^* = 1$. Alice sets $\rho_0 = \gamma$, fully expecting that she will be well treated by Bob. Hence, after period 0 is done, Alice is in the exact situation as when she began, and her incentives are to say "I'd like to revise my plan: I'll set $\rho_1 = \gamma$ and $\rho_2 = 1$, and Bob—if you are evil—you should wait until time 2 to trigger." We can fix things by a similar construction as last paragraph: Letting π_G^0 be the smallest probability that Bob is type-G so that Alice's optimal scheme is $\tau^* = 0$, if $\tau^* = 1$, Alice announces $\rho_0 = \gamma$ and Evil Bob randomizes so that, if he doesn't trigger, Alice assesses probability $1 - \pi_G^1$ that he is evil. And so forth.

Developing these no-commitment equilibria, especially in situations where

Evil Bob comes in multiple flavors, takes another entire paper, so we leave this here (for now), with the following observation: The idea that, without commitment, time-based screening is problematic goes back to Weiss (1983) and Admati and Perry (1987), with subsequent contributions by Noldeke and Van Damme (1990), Swinkels (1999), and Kremer and Skrzypacz (2004). What is novel here is that, to greater or lesser extent, those papers assume the uninformed side of the market is somewhat competitive; here the uniformed side, Alice, leads in specifying terms to Bob. So perhaps the literature that comes closest to the problem raised by a lack of commitment is the literature on the Coase Conjecture, where a monopolist sets prices dynamically to screen amongst prospective buyers.

9. Concluding remarks

The model examined here is, of course, highly stylized, and its precise specifications play a critical role in its analysis. Most significantly and directly, the assumption that both Alice and Bob's payoffs depend linearly in Alice's scale decision makes life relatively simple.

On the other hand, we've assumed that Alice is unsure at the outset about Bob's payoffs, but she knows the consequences of his actions for her. That is, if he treats her well, her payoff is 1, for all types of Bob. And if he treats her poorly, her payoff is $-A$ for all types of Bob, on a scale where non-engagement gives payoff 0. We might suppose instead that different Bobs generate different payoffs for Alice. Since it is natural to assume that Alice "realizes" her payoffs in each stage during that stage, allowing different Bobs to generate different payoffs for her if he treats her well provides her with a source of information that changes the story dramatically. So suppose that if she treated well by Bob, her payoff is 1, regardless of which type of Bob she faces.

We could still imagine that the cost of her of being treated poorly by Bob depends on the type of Bob she faces, especially given our assumption that she never engages with Bob once he treats her poorly. In particular, we could enrich the formulation by assuming that Evil Bob comes in one of N flavors, where the nth Evil Bob gains payoff B_n and inflicts on her cost A_n if and when he treats her poorly.

The analysis, for the most part, doesn't change. Bob's incentives, given $\{\rho_t\}$, remain the same; Alice optimally employs a sequence $\{\rho_t\}$ that has the structure of Proposition 3. Of course, if $N = 1$, nothing at all has changed. And, for $N > 1$, while Alice's optimal choice of $\{\tau_n^*; n = 1, \ldots, N\}$ is a bit more complex, especially if she is optimizing through a numerical search, this choice is only a bit more complex.

References

to be supplied later