Starting Small to Screen for Mr. Good Bob

David Kreps

November 2018

Abstract: A stylized model explores the prescription to "start small" in new and prospectively long-run relationships: Alice, dealing with prospective partner Bob, has a general prior over Bob's intentions and starts small to screen for good types of Bob. The well-known story is that starting small gets undesirable types to reveal themselves at reduced cost, but the analysis raises a second possibility: Start small to encourage undesirable (but patient) types to delay undesirable behavior. Alice does best with undesirable types who are either very impatient or patient; undesirable types whose patience matches her own are most costly to her. (98 words)

Acknowledgements: I'm grateful for helpful comments from Eddie Dekel, Pascaline Dupas, Drew Fudenberg, Faruk Gul, Avner Kreps, Andy Skrzypacz, Joel Sobel, Joel Watson, and participants at seminars given at Tel Aviv University, Hebrew University, Bar Ilan University, and Stanford. The financial support of the Graduate School of Business, Stanford University, is also gratefully acknowledged.

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 times $t = 0, 1, \ldots$ must decide whether to engage with a (prospective) trading partner, Bob. If Alice engages with Bob, Bob can respond by either treating Alice well or poorly. Alice would happily engage with Bob if she knows he will treat her well, but she is uncertain about his motivations for entering into this relationship. 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 identifies whether Bob is a good prospective partner.

One possible screening device is to start small: Alice (if she can) begins by trusting Bob in a small scale engagement, 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 reveal himself early on.

Hence, a micro-finance lender may begin lending a small amount to a new client, with the amount loaned increasing as long as the client pays back prior loans. A manufacturer, dealing with an upstream supplier, develops the scale and importance of its relationship with the supplier, learning how and how well the supplier behaves. A firm employs new employees in low-level jobs, promoting employees in an internal labor market if the employees perform well.

I present and study a highly stylized model of this situation. Alice begins with a fairly general prior assessment over what motivates Bob and uses the time-varying scale of her engagements with Bob to minimize the adverse impact on herself if Bob is, in fact, someone whose incentives are to harm her. We see two different ways in which Alice can employ her ability to scale her engagements: She might engage Bob's *impatience*, getting him to treat her poorly early on, when the scale is small, thus limiting the damage he does to her. And she might engage his *patience*, providing him with incentives to delay treating her poorly while the scale of their engagement grows. The benefit to her in the second case is twofold: She benefits from being treated well while he waits, and discounting blunts the impact on her of his bad behavior when finally he treats her poorly.

The analysis here is limited in several ways:

1. Alice is modeled as a Stackelberg leader with strong powers of commitment. Some justification for this assumption is offered. But 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" (Noldecke and Van Damme, 1990; Swinkels, 1999; Weiss, 1983): After some screening has taken place, Alice might like to reset her behavior; we assume that she can credibly commit at the start not to do so.

2. Alice's uncertainty about Bob concerns (only) his intentions, modeled as his payoffs in the stage-game encounter. In real-life applications, such as those cited previously, the lender / downstream firm / employer is concerned with the borrower's / upstream firm's / employee's (multidimensional) abilities.

3. The encounter between Alice and Bob is a supergame, and the folk-theorem applies; there are many possible equilibria, including equilibria in which Alice, convinced that Bob is motivated to harm her, nonetheless concludes an "agreement" with him in which he is able to harm her *some* of the time, to their mutual benefit. The analysis here considers (only) equilibria in which Alice uses variations of the *grim trigger* strategy, in which she never again trusts Bob, once he treats her poorly.

Despite these limitations, the analysis gives some interesting insights into "starting small" strategies and equilibria.

This idea has been modeled by Sobel (1985), Diamond (1989), Kranton (1996), and especially by Watson (1999a, 1999b).[1] Watson (1999b) is the most

[1] The emphasis here is on learning about Bob's intentions. "Starting small" can also play an incentive role, as in Admati and Perry (1986).

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 many 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 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," hence this paper fits within the general rubric of dynamic mechaism design. (d) Most importantly, in Watson's models (and Sobel's before his), the "evil" type comes in one flavor,[2] 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 flavor 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 Alice receives; she may be worse off with multiple flavors of evilness than she is dealing with each one alone.

2. The base model

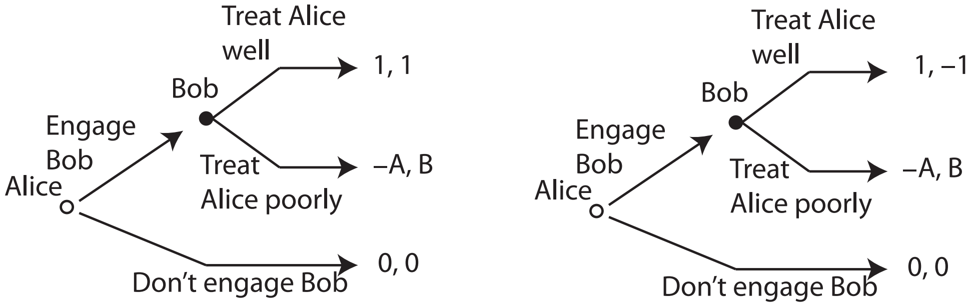
Consider two parties, Alice and Bob, engaged in an infinite-stage game, with stages at times $t = 0, 1, \dots$.

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 $A > 0$. (These are her payoffs for now; they will change as we develop the model.) Her overall payoff for the infinite-stage game is her

[2] Bob may fall into one of several *types*, for instance, the evil type is someone who wishes to be engaged by Alice so he can treat her poorly. And, within each type, Bob may come in several *flavors*; different flavors of Evil Bob differ in their relative payoffs from treating Alice poorly, which in turn determines how impatient they are to do so.

discounted expected sum of payoffs, discounted with discount factor $\delta < 1$.

Alice is unaware of Bob's stage-game payoffs, although she assesses probability 1 that his overall payoffs from the infinite-stage game are the (expected) discounted sum of his payoffs from each stage (with the same discount factor δ), and he has one set of payoffs for all stages. Normalize his payoffs in the stage game so that no engagement gives him payoff 0. Assume that Alice assesses probability 0 that Bob's stage payoff from treating her well equals his stage payoff from no engagement. If he prefers to treat her well over no engagement, normalize his payoff to treating her well to be 1 and say that he is a *type-G* Bob. If he prefers no engagement to treating her well, normalize his payoff to treating her well to be -1 and say that he is a *type-H* Bob. For either type of Bob, this leaves a free parameter, namely his payoff from treating her poorly, which we denote by B. Let π_G be the probability Alice assesses that Bob is type G, F_G be the (conditional) distribution function for her assessment of B, given he is type G, and F_H be the (conditional) distribution function for her assessment of B, given he is type H. See Figure 1.



(a) Type G Bob

(b) Type H Bob

Figure 1. The Basic Stage Game with Two Broad Types of Bob. In both panels, Alice's payoffs are listed first, and $A > 0$. In panel a, Bob is type G (which Alice assesses has probability π_G); her assessment of the free parameter B is given by F_G. In panel b, Bob is type H (which has probability $\pi_H = 1 - \pi_G$); Alice's assessment of B is given by F_H.

It is somewhat natural (in most economic applications) to assume that Bob's payoff from treating Alice well is less than his payoff from treating her well; that is, $B > 1$ for type G and $B > -1$ for type H. But we do not need this assumption and, in particular, for type-G Bobs, a case can be made that the psychological cost of treating Alice poorly outweighs for Bob any tangible costs incurred by treating her well. In any event, we subdivide each of the two main types of Bob into two sub-types:

G1. *Saintly Bob*, a type-G Bob with $B \leq 1$

G2. *Good Bob*, a type-G Bob with $B > 1$

H1. *Bad Bob*, a type-H Bob with $B \leq 0$.

H2. *Evil Bob*, a type-H Bob with $B > 0$.

In words, Saintly Bob prefers treating Alice well to treating her poorly; absent any longer-term considerations, he will do so whenever she engages him.[3] Good Bob, on the other hand, has a short-run incentive to treat her poorly. However, as long as B is not too large—specifically, $B \leq 1/(1-\delta)$—Good Bob will treat Alice well *if treating her poorly means she will not engage with him in the future*. In the literature, Good Bob for $B \leq 1/(1-\delta)$ is Alice's partner in the canonical *Trust* or *Promise Game*. He is called Good Bob here not because he is inherently good, but because he is (as long as $B \leq 1/(1-\delta)$) a good candidate for a long-run partnership with Alice.

On the other hand, Bad Bob is, as long as $B > -1/(1-\delta)$, a bad candidate, in the following sense. He prefers most of all that Alice not engage with him (ignoring the knife-edge case that $B = 0$) and, as long as $B > -1/(1-\delta)$, he would be willing to treat her poorly if, by so doing, he could convince her not to engage with him in the future. For cases $-1 > B > -1/(1-\delta)$, Bad Bob appears in the literature in the canonical *Threat Game* (or, if the stage-game is repeated only finitely many times, the Chain-Store Paradox).

Finally, we have Evil Bob. He wants Alice to engage with him, so he can get his best overall stage-game payoff, by treating her poorly.

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. In this paper, we look

[3] In the knife-edge case $B = 1$, we assume he opts for treating her well.

(only) for 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.*

No formal justification for limiting Alice's strategy in this way can be provided; it is certainly possible, for certain parameterizations, to construct (somewhat complex) equililbria that are Pareto superior to an equilibrium where Alice behaves in this fashion. Indeed, even if Alice knows that Bob is evil, this can happen. 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 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 $B = -1/(1-\delta)$.)
- 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

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

This is (weakly) positive if

$$\phi_\delta \geq \frac{(1-\delta)A}{(1-\delta)A+1}. \tag{1}$$

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 $\pi_G > A/(A+1)$.

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 stage-game 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 $(-\infty, 1/(1-\delta))$.[4]

Having eliminated all the Bad Bobs with this cheap talk, let $\pi_H = 1 - \pi_G$ is the prior probability that Bob is Evil Bob. And, also to simplify the analysis, assume 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 $0 <$ B$_1 < B_2 < \ldots < B_N$, and ϕ_n is the (conditional) probability that Bob, if evil, has paramenter B_n. Call B_n a *flavor* of Evil Bob, and abbreviate Evil Bob with flavor B_n by B_n-EB.

3. *Screening out Evil Bob by starting small*

Evil Bob can be screened out by several different means. For instance,

[4] The treatment of Good Bobs with $B > 1/(1-\delta)$ is dealt with in Section 9.

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 but compensates her sufficiently so that she is happy.[5]

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. With this normalization, if Alice chooses not to engage Bob, his payoff is unaffected. It is Bob's relative payoffs from treating Alice well or poorly that are scaled. It is reasonable to think that Bob's costs and/or benefits from either possibility are monotonic functions of the scale of Alice's engagement with him, but I am assuming much more, namely that they are scaled in equal proportions and, even more, that, the same proportionality applies to Alice's payoffs from engaging him and either being treated well or poorly.[6] This particular parameterization will make the algebra relatively simple. It would be of interest to see what happens if, say, Bob's payff from treating Alice poorly reduces at a slower rate then his payoff from 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. But we do none of that here.

[5] In the context where A loans B money on a short-term basis, A could insist that B post collateral the potential loss of which would cause even Evil Bob to repay his loan. A reason this story is particularly germane to micro-finance is that, in such cases, the borrower may not have sufficient collateral to post.

[6] For reasons that will become clear, I can get by with somewhat weaker assumptions about the payoffs to type-G Bobs, namely that type-G Bob's payoffs scale monotonically in the scale of engagement, and the scale factor of treating Alice poorly is no greater than the scale factor of treating her well.

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 up such a contract. A better story, perhaps, is that Alice has this encounter with many Bobs through time—she is an employer, employing many Bobs, say, or a micro-finance lender dealing with many different borrowers in different stages of their relationship with her—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 at some point, even if $B < 1/(1-\delta)$, if the sequence $\{\rho_t\}$ decreases at some point. 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. (By slightly perturbing the sequence of scales, she can make any response that is best for her a strict best response for him.)

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, anticipating this, would wait until time 1. Alice's optimization problem, then, is to find the best 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.

Recall that Alice's assessment over the different possible flavors of Evil Bob are B_1 through B_N, with $0 < B_1 < B_2 < \ldots B_N$. Let

$$\beta_n := \frac{B_n}{B_n + 1} \quad \text{and} \quad \gamma_n := \delta\beta_n = \frac{\delta B_n}{B_{n+1}}.$$

Note that (of course), if $n > n'$, then $\beta_n > \beta_{n'}$ and $\gamma_n > \gamma_{n'}$.

Lemma 1. *Suppose Alice announces $\{\rho_s\}$. In comparing whether triggering at t or $t+1$ is better, B_n-EB (Evil Bob of flavor B_n) strictly prefers triggering at t if $\rho_t > \gamma_n \rho_{t+1}$, he is indifferent if the two are equal, and he strictly prefers $t+1$ if $\rho_t < \gamma_n \rho_{t+1}$.*

Hence, if for two times t' and t'', with $t' < t''$, Alice's announcement satisfies $\rho_t = \rho_{t''}\gamma_n^{t''-t}$ for $t = t', t'+1, \ldots, t'',$: B_n-EB is indifferent between triggering at any time t between t' and t'': B_n'-EB strictly prefers to trigger at time t' to any time $t'+1, \ldots, t''$, for $n' < n$; and B_n''-EB strictly prefers t'' to any time from t' to t'', if $n'' > n$.

Proof. This is simple algebra. If B_n-EB triggers at time t, his payoff is $\sum_{s=1}^{t-1} \delta^s \rho_s(-1) + \delta^t \rho_t B_n$. Triggering at time $t+1$ yields $\sum_{s=1}^{t} \delta^s \rho_s(-1) + delta^{t+1} \rho_{t+1} B_n$. The difference between these two is

$$\left[\delta^t \rho_t B_n\right] - \left[\delta^{t+1} \rho_{t+1} B_n - \delta^t \rho_t\right] = \delta^t \left[\rho_t B_n + \rho_t - \delta \rho_{t+1} B_n\right],$$

which is strictly positive if and only if $\rho_t(B_n + 1) > \delta \rho_{t+1} B_n$, or $\rho_t > \rho_{t+1}\gamma_n$. The difference is 0 if $\rho_t = \rho_{t+1}\gamma_n$, and it is strictly negative if $\rho_t < \rho_{t+1}\gamma_n$. 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. Set $\overline{\rho} :=$ sup$\{\rho_t; t = 0, 1, \ldots\}$ and let τ be the earliest time such that $\rho_t > \delta\overline{\rho}$. Then for each n, B_n-EB has a finite solution to his optimization problem of when to trigger, which is no later than time τ.*

b. *Suppose that Alice announces $\{\rho_t\}$ with $\rho_t > 0$ for some t. Suppose B_n-EB has, as an optimal solution, to trigger at τ_n (not precluding the possibility that he has several solutions). Then for $n' < n$, all of $B_{n'}$-EB's optimal triggering times are at time τ_n or earlier, and for $n'' > n$, all of $B_{n''}$-EB's optimal triggering times are at time τ_n or later. Moreover, Good Bob of flavor B will never trigger Alice at a time before τ_n, for any B.*

c. *Saintly Bob always treats Alice well, regardless of Alice's announcement. And 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.*

Proof. (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 B_n-EB's payoff from triggering at time τ versus triggering at any time $t > \tau$ is

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

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}. Hence B_n-EB always prefers to trigger at time τ than any time $t > \tau$. And since this puts a finite bound on the number of times he might consider, some one of those times (from 0 to τ) must be best.

(b) If Alice announces $\{\rho_t\}$ with some strictly positive ρ_t, B_n-EB'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 B_n-EB. Then triggering at t must be at least as good for him as triggering at any time $t' > t$, which is

$$\delta^t\rho_t B_n - \sum_{s=0}^{t-1}\delta^s\rho_s \geq \delta^{t'}\rho_{t'}B_n - \sum_{s=0}^{t'-1}\delta^s\rho_s, \quad \text{or} \quad \left[\delta^t\rho_t - \delta^{t'}\rho_{t'}\right]B_n \geq -\sum_{s=t}^{t'-1}\delta^s\rho_s.$$

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 $n' < n$,

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

if $\delta^t\rho_t - \delta^{t'}\rho_{t'} \geq 0$, then multiplying by $B_{n'} > 0$ 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 $\left[\delta^t\rho_t - \delta^{t'}\rho_{t'}\right]B_{n'} > \left[\delta^t\rho_t - \delta^{t'}\rho_{t'}\right]B_n$, and we have the desired strict inequality. In either case, $B_{n'}$-EB for $n' < n$ strictly prefers triggering at t to any subsequent time, and so his optimal time to trigger must be t or less.

The second half of this part of (b), that if $n'' > n$, then optimal times for $B_{n''}$-EB are no smaller than any optimal time for B_n-EB, is just the contrapositive of the first half.

And if t is optimal for B_n-EB, then it must be that $\delta^t \rho_t B_n > \delta^{t'} \rho_{t'} B_{n'}$ for all $t' < t$: For t to be optimal, ρ_t must be strictly positive, and either $\rho_{t'} = 0$, giving the strict inequality, or $\rho_{t'} > 0$, in which case we get a strict inequality because, otherwise, triggering at t' instead of t would give at least as good a reward from triggering, without incurring the strictly positive cost of "waiting" at time t'. But then, for all finite and positive B, $\delta^t \rho_t B > \delta_{t'} \rho_{t'} B$. A Good Bob of flavor B does better triggering at t than at t', since he does strictly better when he triggers (in present value terms) and possibly earns some further positive rewards from treating Alice well while he waits. This is true for any $t' < t$, so any flavor of Good Bob will trigger, if at all, no earlier than time t.

(c) Since not treating Alice well means she never engages again, and since Saintly Bob prefers treating her well in terms of immediate payoff, it is clear that Saintly Bob will always treat her well. As for Good Bob: 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 $\overline{B} = B_N + 1$ and let $\overline{\gamma} = \delta\overline{B}/(\overline{B}+1)$. Suppose Alice announces $\{\rho_t\}$ given by $\rho_t = 1$ for $t \geq T$ and $\rho_t = \overline{\gamma}^{T-t}$ for $t < T$, for some (presumably large) T. Then for all t, $\rho_t/\rho_{t+1} = \overline{\gamma} > \gamma_n$ for all $n = 1, \ldots, N$. Hence, by Lemma 1, all flavors of Evil Bob optimally 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

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

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* $\overline{\gamma}^T$. Even ignoring the summation term (which is positive), since $\delta > \overline{\gamma} = \delta\overline{B}/(\overline{B}+1)$, the term $\pi_G\delta^T/(1-\delta)$ goes to zero more slowly than does $\pi_H\overline{\gamma}^T A$ (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. ■

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 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 B_n-EB is indifferent between triggering at times t and t', for $t > t'$ and such that $\rho_{t'} > 0$. If $A > B_n$, Alice strictly prefers that Bob trigger at time t'. If $A < B_n$, she strictly prefers that he choose t. If $A = B_n$, she is indifferent.*

Proof of Proposition 2. If B_n-EB is indifferent between triggering at times t and t', with $t > t'$, then

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

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 $A > B_n$,

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

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

This says that if $A > B_n$, Alice prefers B_n-EB to trigger sooner rather than later, in terms of his contribution to her overall expected payoff. 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 $B_n > A$, the reverse inequality will hold. ■

5. The form of Alice's optimal announcement

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:*

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

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 $B_n > A$. And since this sequence of scales is nondecreasing , type-G Bobs never trigger.

Figure 2 illustrates this proposition. Suppose that $N = 3$, $\delta = 0.9$, $B_1 = 3$, $B_2 = 10$, and $B_3 = 100$. Then $\gamma_1 = 0.675$, $\gamma_2 = 0.8182\ldots$, and $\gamma_3 = 0.89108910\ldots$ Suppose Alice chooses $\tau_1 = 8, \tau_2 = 12$, and $\tau_3 = 15$. Then $\rho_{15} = 1$; in fact (but not depicted) $\rho_t = 1$ for all $t \geq 15$. Between times 15 and 12 (reading backwards), scales shrink at rate γ_3 per period. Between times 12 and 8, they shrink more rapidly, at rate γ_2. And between times 8 and 0, they shrink most rapidly, at rate γ_1. In Figure 2, the "kinks" in the sequence of scales are fairly apparent. Now suppose $A = 7$. B_1-EB is happy to trigger at times 0 to 8; since $B_1 < A$, he chooses time 0 on Alice's behalf. B_2-EB is happy to trigger at times 8 to 10; since $B_2 > A$, he triggers at time 10. And $B_3 - EB$ triggers at time 15.

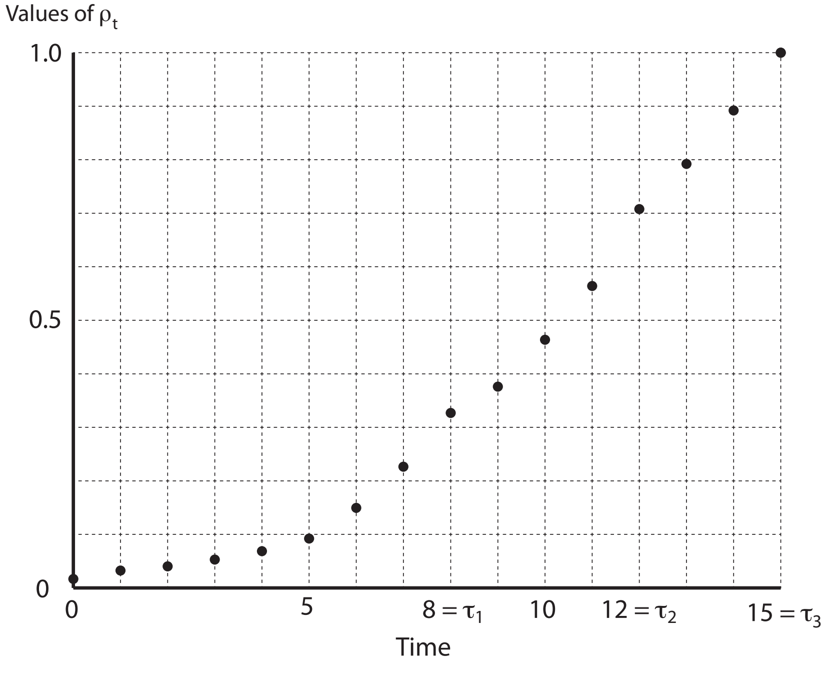


Figure 2. The shape of an "optimal announcement" by Alice. The ρ_t's "descend" geometrically from the value 1 at $\tau_3 = 15$ until $\tau_2 = 12$ at the rate γ_3, then they descend at the higher rate γ_2 until $\tau_1 = 8$, and then at rate γ_1 until $\tau_0 = 0$. Hence B_1-EB sees triggering at any 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.

Several remarks about this proposition are in order.

1. Most importantly, this characterizes Alice's optimal announcement given the sequence $\{\tau_n\}$, but she still must find optimal value of those times. This is not a trivial problem in general, although for $N = 1$, we can almost solve her problem; see Section 6. And this characterizes *an* optimal announcement. In knife-edge cases, she may have more than one.

2. $\tau_n = \tau_{n+1}$ is allowed. Imagine, for instance, that $N = 3$, $B_n < A$ for all three n, and $0 = \tau_0 = \tau_1 < \tau_2 < \tau_3$. Then B_1-EB triggers at time $\tau_0 = \tau_1$, where B_2-EB triggers. Pooling in this sense is certainly a possibility. (Indeed, for π_G close enough to 1, it is clear that Alice will choose $\tau_0 = \ldots = \tau_N = 0$; the expected cost of Evil Bob triggering is so small that she prefers not to lose any value from type-G Bob; she starts with $\rho_0 = 1$.)

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, $\rho_{\tau_{N-1}} = (\gamma_N)^{\tau_N - \tau_{N-1}}$, from which ρ_t for τ_{N-2} to τ_{N-1} is computed, and so forth.

4. In the example in Figure 2, note that the different flavors of Evil Bob trigger at times $\tau_0 = 0$, $\tau_2 = 10$, and $\tau_3 = 15$. There is a "kink" in the sequence of ρ_t at $\tau_1 = 8$, but because $B_1 < A < B_2$, no flavor of Evil Bob triggers at time τ_1. Nonetheless, examples exist showing that, in such cases, $0 < \tau_1 < \tau_2$ can be optimal for Alice; "kinks" of this sort, where no flavor of Evil Bob triggers, are possible.

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 τ_n+1 to triggering at t_n^*.[7] Graphically, the situation must be as depicted in Figure 3, where the two curves are the functions ψ_n and ψ_{n+1}, and solid dots represent values of ρ_t.

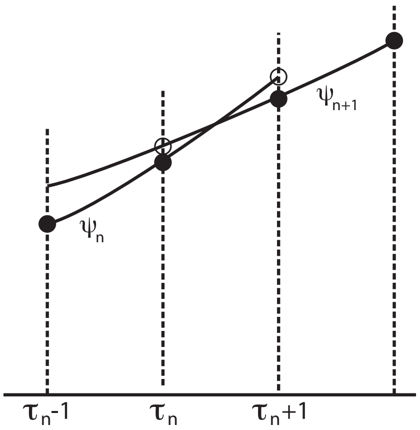


Figure 3. The hypothesized situation around τ_n. See text for explanation.

But this situation is inconsistent with Alice optimizing except for a knife-

[7] 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.

edge case and, in that knife-edge case, we can adjust the ρ_ts so that τ_n is 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 no further 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 τ_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 flavor of Evil Bob*[8]

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, I can provide Alice with considerable assistance, as well as give several comparative statics results. First, I 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 nonnegative integer 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 to do), while type-G Bob always treats Alice well. Then Alice's expected payoff, which we write $\mathcal{R}(\tau; \delta, A, B, \pi_G)$, is

$$\pi_G \left[\sum_{s=0}^{\tau-1} \delta^s \gamma^{\tau-s} + \delta^\tau \frac{1}{1-\delta} \right] - \pi_H \gamma^\tau A = \pi_G \delta^\tau \left[\beta \frac{1-\beta^\tau}{1-\beta} + \frac{1}{1-\delta} \right] - \pi_H \gamma^\tau A. \quad (2)$$

Recall that, in Section 3, when introducing the reason for starting small, 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.” We see precisely this strategy in the right hand side of Equation 2. By choosing $\tau > 0$, Alice reduces the impact

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

of Evil Bob's (immediate) trigger; this, or rather the expectation of this, is the final term $-\pi_H \gamma^\tau A$. But this comes as a cost in terms of what she accrues from Bob if he is type G; this is the first term in the final expression.

However, when $B > A$, Alice's strategy is based on a different idea. She entices Evil Bob to wait until time τ to trigger. Her expected payoff is

$$\begin{aligned}\mathcal{L}(\tau; \delta, A, B, \pi_G) : &= \sum_{s=0}^{\tau-1} \delta^s \gamma^{\tau-s} + \delta^\tau \left[\pi_G \frac{1}{1-\delta} - \pi_H A \right] \\ &= \pi_G \delta^\tau \left[\beta \frac{1-\beta^\tau}{1-\beta} + \frac{1}{1-\delta} \right] + \pi_H \delta^\tau \left[\beta \frac{1-\beta^\tau}{1-\beta} - A \right]. \qquad (3)\end{aligned}$$

In this case, she benefits if Bob is evil in two ways (in the second term on the r.h.s. of Equation 3): He treats her well until time τ, which is good for her. And the delay in when he triggers reduces its impact by the discount factor δ^τ. Traded off against these benefits is the reduction in what she accrues from Bob if he is type G, which is the first term on the r.h.s. of Equation 3.

The formulas in Equations 2 and 3 can be used to identify the optimal value of τ up to one of two adjacent integers. To do this, first extend the domain of definition of the functions in Equations 2 and 3 from integer τ to $\xi \in (-\infty, \infty)$. Following some algebraic manipulation, we get

$$\mathcal{R}(\xi; \delta, A, B, \pi_G) = \pi_G \delta^\xi \left[B + \frac{1}{1-\delta} \right] + \gamma^\xi \left[\pi_G (A - B) - A \right], \quad \text{and} \qquad (2')$$

$$\mathcal{L}(\xi; \delta, A, B, \pi_G) = \delta^\xi \left[B + \frac{\pi_G}{1-\delta} - (1 - \pi_G) A \right] - \gamma^\xi B. \qquad (3')$$

Proposition 4. *For fixed δ, A, B, and π_G, both $\mathcal{R}$ and $\mathcal{L}$ are single-peaked in ξ. $\mathcal{R}$ achieves its maximum at*

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

while, if $B > A$, $\mathcal{L}$ achieves its maximum at

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

Hence, if $A \geq B$, the optimal value of τ for Alice is 0 if $\xi^*_{\mathcal{R}} \leq 0$ and is either $\lfloor \xi^*_{\mathcal{R}} \rfloor$ or $\lfloor \xi^*_{\mathcal{R}} \rfloor + 1$ otherwise; if $B > A$, the optimal value of τ for Alice is zero if $\xi^*_{\mathcal{L}} < 0$ and either $\lfloor \xi^*_{\mathcal{L}} \rfloor$ or $\lfloor \xi^*_{\mathcal{L}} \rfloor + 1$ otherwise.

Proof. The derivative of $\mathcal{R}$ in ξ is

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

and so

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

The sign of $\mathcal{R}'$ is the same as the sign of $\mathcal{R}'/\delta^\xi$, of course, and the latter is the sum of a strictly negative constant (negative, since $\ln(\delta) < 0$) and strictly positive term (since both $\ln(\gamma)$ and $\pi_G(A-B) - A$ are strictly negative) that is continuous and strictly decreasing in ξ, unbounded above as $\xi \to -\infty$ and with limit zero as $\xi \to \infty$. Hence $\mathcal{R}'$ is zero at the solution (in ξ) of

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

$\mathcal{R}'$ is strictly positive for ξ less than this solution and strictly negative for ξ greater than this solution. And simple algebra shows that the solution is $\xi^*_{\mathcal{R}}$ as in the statement of the proposition.

And, for $B > A$, the derivative of $\mathcal{L}$ in ξ is

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

and so

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

The sign of $\mathcal{L}'$ and $\mathcal{L}'/\delta^\xi$ are the same, if $B > A$ (so the term in square brackets is definitely positive), $\mathcal{L}'/\delta^\xi$ is the sum of a strictly negative term that is constant in ξ and a strictly positive term (since $-\ln(\gamma) > 0$) that is continuous and strictly decreasing in ξ, with limits $+\infty$ as $\xi \to -\infty$, and 0 as $\xi \to \infty$. Hence $\mathcal{L}'$ is strictly positive up to $\xi^*_{\mathcal{L}}$ and strictly negative afterwards, where $\xi^*_{\mathcal{L}}$ is the solution to

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

Simple algebra shows that the solution is given by the formula for $\xi^*_{\mathcal{L}}$ provided in the statement of proporition.

The rest of the proposition follows because the functions $\mathcal{R}$ and $\mathcal{L}$ as single-peaked, increasing up to the points $\xi^*_{\mathcal{R}}$ and $\xi^*_{\mathcal{L}}$, respectively, and decreasing thereafter. ■

I turn next to comparative-statics results. For $N = 1$, the situation is described by the four parameters $\delta, A, B,$ and π_G. From these are derived the optimal value of τ, which we denote τ^*, Alice's expected overall payoff, which we denote $\mathcal{V}$, and the overall payoffs for type-G Bobs, denoted $\mathcal{U}_G$, and for Evil Bob, denoted $\mathcal{U}_{EB}$. (Note that in Alice's optimal scheme, Good and Saintly Bobs of any flavor get the same deterministic payoff; Evil Bob's payoff is likewise deterministic. Hence the adjective "expected" is relevant only for Alice.) In general, we write these endogenous values as functions of the exogenous parameter(s) of interest: $\mathcal{V}(\delta, A, B, \pi_G)$ denotes Alice's expected payoff as a function of all four parameters, while $\mathcal{V}(B)$ (say) denotes her expected payoff as a function of B, with the other three parameters held fixed.

Proposition 5. *Assume that $\pi_G > 0$ throughout.*

a. $\mathcal{V}$ is strictly increasing in δ and π_G. It is strictly decreasing in A.

b. $\mathcal{V}$ is V-shaped in B. It is nonincreasing in B for $B < A$, and it is nondecreasing in B for $B > A$.

c. τ^* is nonincreasing in π_G and nondecreasing in A. Hence, $\mathcal{U}_G$ and $\mathcal{U}_{EB}$ are nondecreasing in π_G and nonincreasing in A.

This leaves as open the behavior of τ^*, $\mathcal{U}_G$, and $\mathcal{U}_{EB}$ in δ and B. Based on a variety of numerical examples, I conjecture that τ^* is nondecreasing in both δ and B and that $\mathcal{U}_G$ is nondecreasing in δ and nonincreasing in B. (If τ^* is nondecreasing in B, then $\mathcal{U}_{EB}$ must be nonincreasing in the same parameter.) After proving Proposition 5, I'll discuss these conjectures.

Proof of Proposition 5(a). 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. For fixed δ and B, there is a finite upper limit to the possible triggering times for Evil Bob, hence there is a finite number of candidates for τ. Apply the standard result that the upper envelope of a finite number of strictly increasing [resp., decreasing] functions is strictly increasing [resp., decreasing], to conclude that $\mathcal{V}$ is strictly increasing in π_G and strictly decreasing in A.

To show monotonicity of $\mathcal{V}$ in δ: For $A \geq B$, rewrite Equation 2 as

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

Where this is positive, it is clearly strictly increasing in δ. And we know that it is strictly positive for very large τ, hence it is strictly increasing for all τ's that are candidates for τ^*. Hence $\mathcal{V}$, which is the upper envelope of $\mathcal{R}(\tau)$, is strictly increasing by the usual argument.

And if $B > A$, rewrite Equation 3 as

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

This is clearly strictly increasing in δ where it is positive, and (since $B > A$) is it clearly strictly positive for some τ, so the usual argument once again applies. ■

Perhaps the most interesting part of Proposition 5 is part b, that $\mathcal{V}$ is V-shaped in B, with the kink at $B = A$. To make this vivid, Figure 4 depicts the behavior of $\mathcal{V}$ as B varies, for the case $\delta = 0.9, A = 8$, and $\pi_G = 0.35$.

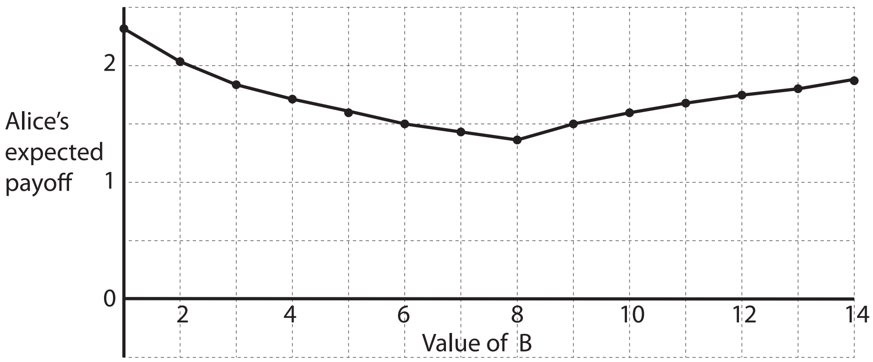


Figure 4. *$\mathcal{V}$ for varying B, for the parameters $\delta = 0.9$, $A = 8$, and $\pi_G = 0.35$.*

One can show that $\mathcal{R}$ is nonincreasing in B for fixed τ (for $A > B$) and $\mathcal{L}$ is nondecreasing (for $B > A$) and apply the usual argument, but here is a direct proof that conveys the intuition behind part b:

Proof of Proposition 5(b). Fix $\hat{B} < A$. Let $\hat{\tau}*$ be the optimal value for τ for $\hat{B}$. Of course, $\mathcal{V}(\hat{B})$ is based on $\hat{B}$-EB triggering at time 0, at the scale $\hat{\rho}_0 = (\gamma_{\hat{B}})^{\hat{\tau}^*}$. Consider, if instead of $\hat{B}$, Evil Bob's payoff from triggering is $\check{B} < \hat{B}$, and consider $\check{\rho}_t = \min\{1, \hat{\rho}_0/\gamma^t_{\check{B}}\}$. In words, Alice begins with $\hat{\rho}_0$ but increases the scales at the rate appropriate for $\check{B}$-EB. Of course, this sequence $\{\check{\rho}_t\}$ causes $\check{B}$-EB to be indifferent between triggering at time 0 and all times up until the time that $\hat{\rho}_0/\gamma^t_{\check{B}} > 1$. Since $\check{B} < \hat{B} < A$, Alice prefers that, faced with this sequence of scales, $\check{B}$-EB trigger at time 0, which damages her just as much as if she were facing $\hat{B}$-EB with the optimal sequence for $\hat{B}$-EB. And, subsequent to time 0, this gives scales ρ_t that are at least as large as those that are optimal for $\hat{B}$-EB (and strictly larger for some t as long as $\hat{\tau}^* \geq 2$). Hence, having eliminated $\check{B}$-EB at time 0, this gives Alice at least as much (indeed, more, if $\hat{\tau}^* \geq 2$) as she would get facing $\hat{B}$-EB, using her optimal strategy against $\hat{B}$-EB. Her optimal strategy against $\check{B}$-EB can do no worse (and, in fact, even better, if $\hat{\tau}^* \geq 2$), hence $\mathcal{V}(\check{B}) \geq \mathcal{V}(\hat{B})$.

Fix $\hat{B} > A$. Let $\hat{\tau}*$ be the optimal value for τ for $\hat{B}$. Of course, $\mathcal{V}(\hat{B})$ is based on $\hat{B}$-EB triggering at time $\hat{\tau}^*$, at the scale 1. Consider, if instead of $\hat{B}$, Evil Bob's payoff from triggering is $\check{B} > \hat{B}$, and consider $\check{\rho}_t = \min\{1, (\gamma_{\check{B}})^{\hat{\tau}^* - t}\}$. In words, Alice, facing $\check{B}$-EB, is using scales appropriate for $\check{B}$-EB, together with $\hat{\tau}^*$. $\check{B}$-EB triggers at time $\hat{\tau}^*$, and so the damage done to Alice is the same as the damange done to her by $\hat{B}$-EB. But up to time $\hat{\tau}^*$, the scales $\check{\rho}_t$ are all at least as large as those under $\hat{\rho}_t$ (and strictly greater if $\hat{\tau}^* \geq 1$), so Alice, using $\{\check{\rho}_t\}$, is better off (strictly, if $\hat{\tau}^* \geq 1$) against $\check{B}$-EB than she is (optimally) against $\hat{B}$-EB. The optimal scheme for her against $\check{B}$-EB can only improve matters for her, hence $\mathcal{V}(\check{B}) \geq \mathcal{V}(\hat{B})$. ■

Do not get lost in all the hats and checks; this argument is really quite intuitive. For B's less than A, Alice's strategy (and Evil Bob's response) plays on the relative *impatience* of Evil Bob to get his reward B. The smaller is B, the more he is (relatively) impatient, so if Alice starts with the same initial scale ρ_0, she can increase it faster for smaller B and keep Evil Bob willing to trigger immediately. Essentially, she is using his impatience against him; smaller B means he is more impatient, relatively speaking, so she does better with this strategy, which plays on his impatience.

But for B's greater than A, Alice is using Evil Bob's *patience* against him. He is willing to wait for his big payoff and, the larger is B, the greater is his willingness to wait. In the proof, Alice is keeping the time τ that Evil Bob triggers the same. And a more patient (larger B) Evil Bob can be kept happy with scales before time τ that rise more slowly.

Proof of Proposition 5(c). 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).

Finally, it is clear that all types of Bob dislike increases in τ^*, everything else held equal. Changes in A or π_G only affect Bob through τ^*, so the second half of part c follows immediately from the first half. ■

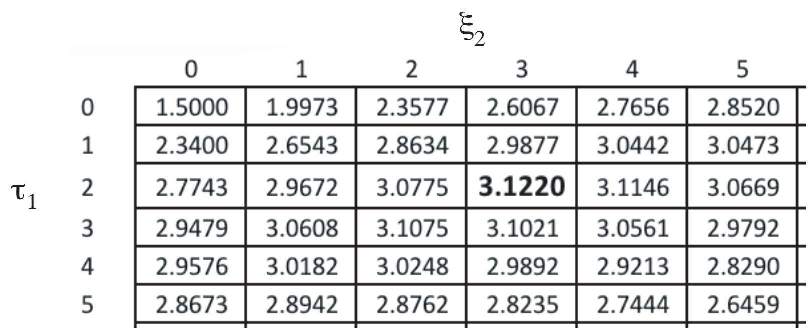
As noted previously, this leaves open the questions of how τ^*, $\mathcal{U}_G$, and $\mathcal{U}_{EB}$ vary with δ and B. I strongly suspect that τ^* is nondecreasing in B and in δ. The first, if true would imply that $\mathcal{U}_G$ is nonincreasing in B. But, even if my conjectures about τ^*'s behavior in B and δ are correct, the behavior of $\mathcal{U}_G$ in δ and $\mathcal{U}_{EB}$ in both B and δ would not be monotonic, because I have formulated Alice's problem as one of choosing integer τ. Consider, for instance, the behavior of $\mathcal{U}_G$ in δ. As δ increases, if my conjecture is correct, there would be a value δ' such that $\tau^*(\delta' - \epsilon) = \tau^*(\delta + \epsilon) - 1$ for all small ϵ. That is, at the value δ', τ^* jumps up by one. Now as δ increases, if there is no change in τ^*, type-G Bob sees his payoff increase, because the same period-by-period payoffs are discounted less. But the sudden jump in τ^* gives a discontinuous jump downward in type-G Bob's payoff. Similar effects arise for Evil Bob when either δ or B vary by small amounts over a range where τ^* jumps discontinuously.

The obvious "cure" for this bad behavior would be to reformulate this model in continuous time. I'll address that variation in Section 8.

7. Two (or more) types of Evil Bob

With two or more types of Evil Bob, clear and clean results are difficult to derive. For the case $N = 2$, for instance, Alice's optimal set of scales $\{\rho_t\}$ is determined by Proposition 3 once she determines the optimal values of τ_1 and τ_2. But those optimal values involve complex interactions between the various parameters of the model.

To find Alice's optimal τ_1 and τ_2 for particular parameter values, a brute-force search is, of course, available. Consider Table 1. This gives Alice's expected payoff as a function of τ_1 and $\xi_2 = \tau_2 - \tau_1$ for the following set of parameters:[9] $\delta = 0.9$, $A = 7$, $B_1 = 3$, $\mathrm{B}_2 = 25$, $\pi_G = 0.5$, and $\phi_1 = 0.7$. (Recall that ϕ_1 is the probability that Evil Bob is flavor B_1, conditional on Bob being evil.) The best expected value for Alice over the region searched—$0 \leq \tau_1 \leq 5$ and $0 \leq \xi_2 \leq 5$ —is $\tau_1 = 2$ and $\xi_2 = 3$ (hence $\tau_2 = \tau_1 + \xi_2 = 5$).



		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
τ_1	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 1. 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.

Since the search represented in Table 1 is limited in scope, it is legitimate to ask whether better values are available for Alice, outside this score. In response to this question:

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.

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

[9] The reader may conclude that the reason for replacing τ_2 in the search with $\xi_2 = \tau_2 - \tau_1$ is that this is more convenient for building this table in a spreadsheet. While that is so—it is indeed more convenient—there is another reason, which is provided below.

the case of a single value of B. This is generally true: in the general (finitely many B) case, simple computations show the following: Reframe Alice's problem as one of choosing $\xi_1 = \tau_1 - \tau_0 = \tau_1, \xi_2 = \tau_2 - \tau_1, \ldots, \xi_N = \tau_N - \tau_{N+1}$. 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's 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 $\delta^{\xi_{n'}}$.

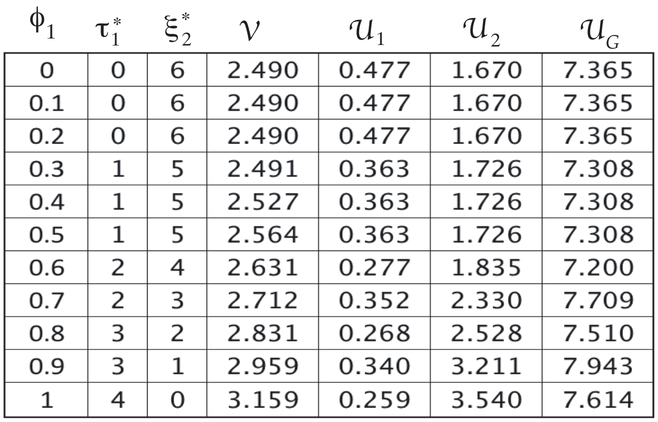
And *if* the $\{\xi_n\}$ that Alice is looking at are such that $K_2 > 0$—which makes sense, since she is presumably looking in regions where her overall expected payoff is positive and, post $\tau_{n'}$, she has shed some of the Evil Bobs that are bad for her—then the proof that $\mathcal{V}$ (viewed 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.

Of course, for discrete τ_n's, this doesn't imply that if, in a search of the sort represented by Table 1, Alice finds specific values of $\tau_1 = \xi_1$ and ξ_2, call them ξ_1^* and ξ_2^*—such that the corresponding $\mathcal{V}$ is a row- and column-maximimum, then this must be Alice's global maximum. But if Alice continues her search a few steps further out from ξ_1^* and ξ_2^* and $\mathcal{V}$ continues to be maximized at those values, this is a strong indication that she can stop searching.

To illustrate the range of phenomena that can occur with $N \geq 2$, Table 2 presents three numerical examples. 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; $\mathcal{U}_1$, the payoff to B_1-EB; $\mathcal{U}_2$, the payoff to B_2-EB; and $\mathcal{U}_G$, the payoff to any type-G Bob. Panels b and c provide the same data for $B_1 = 3 < A = 7 < B_2 = 25$, and $A = 5 < B_1 = 7 < B_2 = 30$, respectively

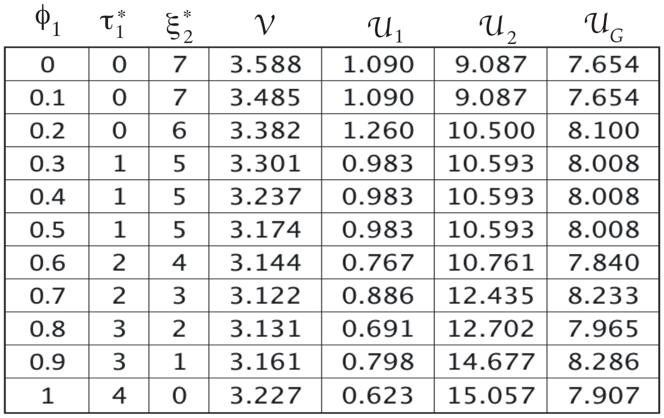
Please note:

- 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.
- Moreover, 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.) Of course, this results from the discrete nature of Alice's options.
- 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 $\phi_1 = 0.1$ 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.



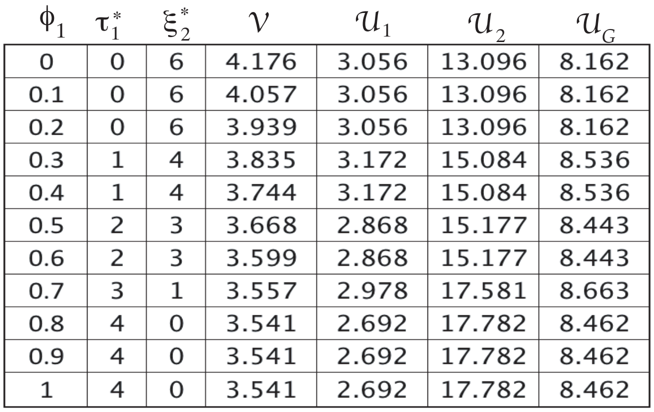
ϕ_1	τ_1^*	ξ_2^*	$\mathcal{V}$	$\mathcal{U}_1$	$\mathcal{U}_2$	$\mathcal{U}_G$
0	0	6	2.490	0.477	1.670	7.365
0.1	0	6	2.490	0.477	1.670	7.365
0.2	0	6	2.490	0.477	1.670	7.365
0.3	1	5	2.491	0.363	1.726	7.308
0.4	1	5	2.527	0.363	1.726	7.308
0.5	1	5	2.564	0.363	1.726	7.308
0.6	2	4	2.631	0.277	1.835	7.200
0.7	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^*	$\mathcal{V}$	$\mathcal{U}_1$	$\mathcal{U}_2$	$\mathcal{U}_G$
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^*	$\mathcal{V}$	$\mathcal{U}_1$	$\mathcal{U}_2$	$\mathcal{U}_G$
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 2. 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 ($\mathcal{U}_1$,$\mathcal{U}_2$, and $\mathcal{U}_G$, 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.

- Note in panel b that Alice's expected payoff is not monotonic in ϕ_1. She is better off dealing with either B_1-EB or B_2-EB than she is dealing with both, for ϕ_1 between 0.5 and 0.9.
- In panel c, Alice loses expected payoff moving from $\phi_1 = 0$ to $\phi_1 = 0.1$, but she does not lose moving from $\phi_1 = 1$ to $\phi_1 = 0.9$ This may at first seem odd, because in each case she is staying with a strategy optimal for one flavor of Evil Bob but is not optimal for the other. The explanation is: When $\phi_1 = 1$, she optimally chooses $\tau_1^* = \tau_2^* = 4$. Since $B_2 > B_1 > A$, both flavors of Evil Bob trigger at time 4. As ϕ_1 decreases to 0.9, her optimal strategy doesn't change, and although B_1-EB gets less than he would if he were alone, this has no impact on Alice. But, on the other side, the optimal values for Alice when $\phi_1 = 0$ and 0.1 are $\tau_1^* = 0$ and $\tau_2^* = 6$. B_2-EB triggers at time 6, while B_1-EB triggers at time 0. The latter is worse for Alice than the former, so as ϕ_1 increases, her expected payoff decreases. To get them both to trigger at time 6 is feasible for her but very suboptimal; to get B_1-EB to wait until time 6, Alice must set scales up to time 6 well below what she can get away with if she only entices B_2-EB to do so.
- 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 $\tau_2^* = \tau_1^* + \xi_1^*$, 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.
- 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,[10] 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. *Continuous time*

Virtually everything we've done here can be done instead in a continuous-time formulation in which, following Alice's announcement of the scales $\{\rho_t; t \in [0, \infty)\}$ that she will use unless and until Bob treats her poorly, Bob (of whichever type) chooses a time T at which he will trigger. Until time T, Bob treats Alice well, with a flow payment at rate 1 (times the scale at each time) accruing to Alice and with Bob getting his corresponding flow payment; then at time T, Alice suffers a discrete loss $-A$ while Bob gains a discrete payoff of B.

A continuous-time formulation does have advantages: Still assuming that there are finitely many flavors of Evil Bob, the calculations of Alice's optimal choices of times τ_n^* becomes an exercise in calculus: For the case $N = 1$, the calculation of τ^* is completely analogous to the computation of $\mathcal{R}'(\xi) = 0$ or $\mathcal{L}'(\xi) = 0$, whichever is appropriate, although the formulas for $\mathcal{R}$ and $\mathcal{L}$ are slightly different. And, more usefully, for general N, if the problem for Alice is formulated in terms of $\xi_1 = \tau_1$, $\xi_2 = \tau_2 - \tau_1$, and so forth, finding the optimal values of the ξ_n involves the simultaneously solution of n partial derivatives. (The uni-modal character of Alice's payoff in each of the ξ_n carries through.) Moreover, this will (probably) mean cleaner comparative statics in the payoffs for types of Bob, since we will no longer have to contend with discontinuous jumps in the τ_n.

Indeed, one can reformulate further with a continuous distribution of the flavors of Evil Bob, although that makes the mechanics of Alice's optimization exercise considerably more complex.

[10] 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.

Notwithstanding these advantages, I've chosen to present this work in a discrete-time formulation because it is easier to comprehend and, in particular, it makes the proof of the main result, Proposition 3, easier.

9. Good Bob with very large B

I assumed that, for type-G Bobs, F_G has support bounded above by $1/(1-\delta)$, so that every flavor of type-G Bob would prefer treating Alice well to treating her poorly, if the former meant ongoing engagement at nondecreasing scale and the latter meant no more engagement. This assumption simplifies the exposition but is unnecessary: Good Bobs with $B > 1/(1-\delta)$, as long as there are finitely many, are "more patient" forms of Evil Bob and are dealt with in similar fashion. That is, if we suppose there are two flavors of Evil Bob with flavors $B_1 < B_2$, and then (say) two flavors of Good Bob with parameters $B_4 > B_3 > 1/(1-\delta)$, then Alice's optimal announcement of scales involves four values of τ_n, where from time 0 to time τ_1, she keeps B_1-EB indifferent, from τ_1 to τ_2, B_2-EB is indifferent, from τ_2 to τ_3, B_3-Good Bob is kept indifferent, and from τ_3 to τ_4, B_4-GB is kept indifferent. Over the periods where flavors of Good Bob are indifferent, the scale of engagement grows at a rate $1/\gamma$ where $\gamma = (1+\delta B)/B$ for the "active" flavor B of Good Bob. Note that if $B > 1/(1-\delta)$, then γ so defined is < 1, so this works.

10. Commitment by Alice?

I'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.

If there is no chance that Bob is saintly—for instance, if Alice believes that Bob of any stripe finds it costly to treat her well versus treating her poorly—a perfect equilibrium is easily constructed in which it is always in Alice's interests to carry out an announced sequence of scales: Imagine that she announces the sequence of scales and then, if she ever deviates, Good Bob "infers" that Alice plans never to engage in the future, hence his best response is to trigger while he can.[11]

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 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.

[11] Of course, no prior announcement is needed; it suffices for the "appropriate scale sequence" to be common knowledge as part of the equilibrium. However, if Alice assesses positive probability that Bob is saintly, this won't work. Even if Saintly Bob believes, after Alice deviates from her announced scales, that she will never engage again, he treats her well. And, subsequent to this, Alice must infer that he is saintly (if Good Bob follows the strategy of always mistreating her), so her sequentially rational response is to engage fully, subsequently. Which changes the optimal strategy for Good Bob, and the entire construction unravels.

- 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 an entire other paper, so I leave this here (for now, at least), with the following observation: The idea that, without commitment, time-based screening is problematic goes back to Weiss (1983), with subsequent contributions by Noldeke and Van Damme (1990), Swinkels (1999), and Kremer and Skrzypacz (2007). Because Alice acts as Stackelberg leader in setting the scale of each engagement, the literature on the Coase Conjecture (e.g., Gul, Sonnenschein, and Wilson, 1986) is also obviously germane.

9. Other variations

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, I'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.

I 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

Admati, Anat R., and Motty Perry (1991), "Joint Projects without Commitment," *Review of Economic Studies*, Vol. 58, 259-76.

Diamond, Douglas W. (1989), "Reputation Acquisition in Debt Markets," *Journal of Political Economy*, Vol. 97, 828-62.

Gul, Faruk, Hugo Sonnenschein, and Robert Wilson (1986), "Foundations of Dynamic Monopoly and the Coase Conjecture," *Journal of Economic Theory*, Vol. 39, 155-90.

Kranton, Rachel (1996), "The Formation of Cooperative Relationships," *Journal of Law, Economics, and Organization*, Vol. 12, 214-33.

Kremer, Ilan, and Andrzej Skrzypacz (2007), "Dynamic Signaling and Market Breakdown," *Journal of Economic Theory*, Vol. 133, 58-82.

Noldecke, Georg, and Eric Van Damme (1990), "Signalling in a Dynamic Labour Market," *Review of Economic Studies*, Vol. 57, 1-23.

Sobel, Joel (1985), "A Theory of Credibility," *Review of Economic Studies*, Vol.

52, 557-73.

Swinkels, Jeroen M. (1999), “Education Signalling with Preemptive Offers,” *Review of Economic Studies*, Vol. 66, 949-70.

Watson, Joel (1999), “Starting Small and Renegotiation,” *Journal of Economic Theory*, Vol. 85, 52-90.

Watson, Joel (2002), “Starting Small and Commitment,” *Games and Economic Behavior*, Vol. 38, 176-99.

Weiss, Andrew (1983), “A Sorting-cum-Learning Model of Education,” *Journal of Political Economy*, Vol. 91, 420-42.