Daniel Altman
7/30/2011
Working Paper

Deterrence Never Works, QED

This paper provides a simple formal model proving that deterrence by punishment never works against a challenger than can ultimately prevail on the battlefield. Regardless of the size of the costs that the deterrer would be able to inflict in response to aggression, a challenger that has the ultimate ability to prevail militarily can always credibly threaten to attack. As a result, it will always be in the deterrer's interest to capitulate in order to avoid the costs of war.

Suppose a state desires a piece of territory from its adversary, and it can seize that territory militarily after a costly war.[1] This adversary aims to deter this attack, and it is able to inflict costs twice the size of the value of the territory to the potential aggressor. Deterrence theory at its simplest predicts that deterrence would succeed in this case, because the costs exceed the benefits. Invasion cannot be threatened credibly and would not occur if the actors are rational and fully informed. This prediction, however, relies on an implicit and implausible assumption that war is an all-or-nothing enterprise. Without that assumption, deterrence no longer works as expected.

Because the costs of war are twice the benefits, the challenger cannot credible threaten to seize the disputed land by force. However, suppose that this challenger has the option of fighting half the war, choosing strategies anew at this decision-point, and then finishing if it wishes. Now suppose the challenger has fought to this halfway point. At this point, the costs of finishing the war (half the costs of fighting the war) no longer exceed the benefits. The challenger can credibly threaten to finish the war from halfway. The deterrer is better off capitulating than fighting a costly war only to lose anyway, so it too would back down rather than fight the second half of the war. In short, the challenger need only get halfway to win.

Can the challenger threaten to fight the first half of the war? If it makes it halfway, from that point it can expect the deterrer to capitulate. Therefore, its choice is no longer to incur costs twice the benefits of victory, but rather merely the costs of fighting the first half of the war. These costs are no larger than the benefits, and so the challenger can credibly threaten to make it halfway. Knowing this, it is again in the deterrer's interest to capitulate rather than endure the costs of half a war only to lose anyway. Overall, the result is that the deterrer capitulates up front despite its ability to inflict costs on the aggressor equal to twice the value of victory.

Deterrence is defeated by a specific strategy premised on exploiting sunk costs. Challengers leverage their ability to credibly threaten to finish a war once they have sunk the costs leading up to that point in order to convince the deterrer to capitulate at the outset.

---

[1] This assumption requires a brute force option for the challenger, in addition to a coercive option. Thomas C. Schelling, *Arms and Influence* (New Haven: Yale University Press, 1967), pp. 2-6.

The underlying intuition is that aggressors facing rational adversaries need not plan to fight through to a military conclusion; they must only fight until the deterrer's interests lead it to capitulate. This cripples deterrence.

This logic extends to costs of any size relative to the benefits of seizing the stakes, not just costs twice the benefits as in the simplified case above. The logic is a direct extension. The challenge can always credibly threaten to finish the last stage of the war in which the costs no longer exceed the benefits. In effect, this final stage disappears, because the deterrer would capitulate. This need not happen only once. In theory, it can happen infinite times (Figure 3). For each new "last" stage, the deterrer cannot credibly threaten to fight, so that fraction of the costs of war need not be borne by the challenger and that stage of the war effectively disappears. Eventually, the remaining costs fall below the benefits of victory, and the challenger can credibly threaten to attack. As a result, the deterrer capitulates at the outset.

As the model is adjusted towards allowing instant-by-instant decisions rather than one all-or-nothing decision between war and peace, the costs that the deterrer must threaten go to infinity and deterrence falls apart. Although a number of the assumptions built into this extremely simple model are open to challenge, this particular change would seem to relax an unrealistic assumption in existing models. Aggressors presumably do have the option of fighting part of a war and deciding whether to continue at any point in a war, so this change to the standard deterrence model would seem to add realism even as its equilibrium outcome seems to strain credulity.

Players
1: Deterrer
2: Challenger

Strategies for Player 1
C: Capitulate
R: Resist

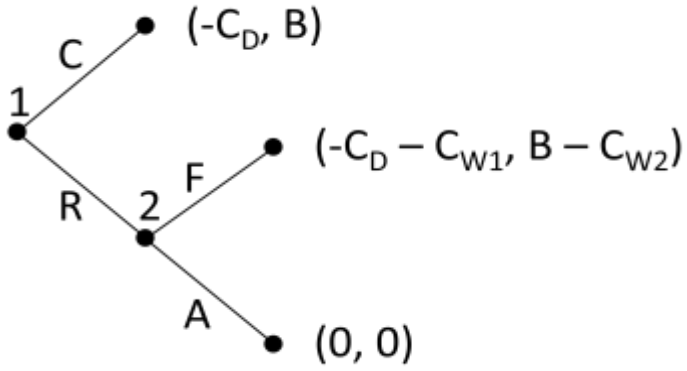
Strategies for Player 2
A: Abandon challenge
p: The proportion of the war to fight
p_1, p_2, …, p_n: the proportion of the war to fight at decision-points 1,2, …, n
n: the number of decision-points

Elements of the Payoffs
B: Benefits of achieving the prize for the challenger
C_D: Costs of defeat (losing the prize) for the deterrer
C_{W1}: Costs of fighting the war for the deterrer
C_{W2}: Costs of fighting the war for the challenger

p: The proportion of the war fought

$p_1, p_2, \ldots, p_n$: the proportion of the war fought after decision-points 1,2, …, n

Figure 1: Standard Deterrence Theory[2]

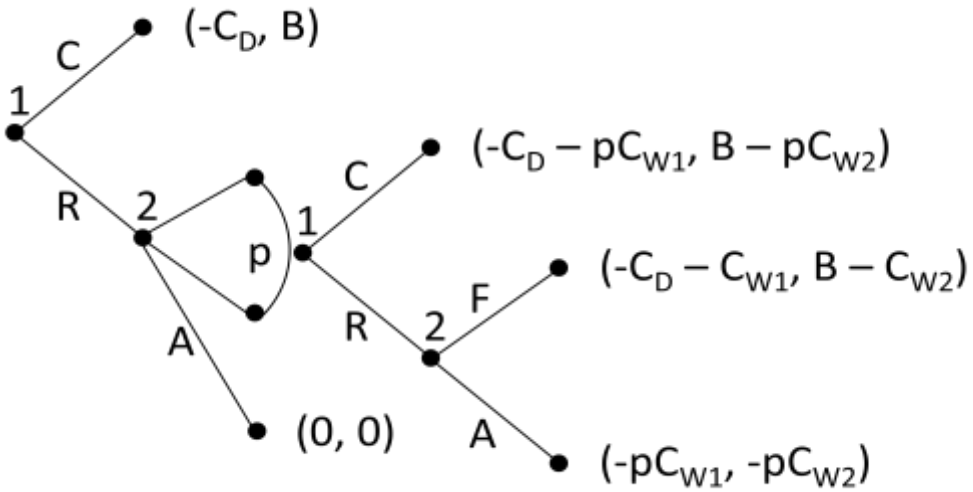


Equilibria Outcomes:

$(-C_D, B)$ if $C_{W2} \leq B$ (Deterrer Capitulates)

$(0, 0)$ if $C_{W2} > B$ (Challenger Abandons the Challenge)

Figure 2: One Decision-Point



Solution

Challenger chooses F in last stage if

$B - C_{W2} \geq -pC_{W2}$

$B \geq (1-p)C_{W2}$

Challenger chooses p if

$B - pC_{W2} \geq 0$

[2] I follow Schelling (1966) in placing the burden of the last move on the challenger in deterrence scenarios. Schelling, *Arms and Influence*.

$B \geq pC_{W2}$

Combining these, the deterrer capitulates in the first stage if there is some p meeting these conditions

$B \geq C_{W2} - pC_{W2}$

$p \geq 1 - B/C_{W2}$

And

$p \leq B/C_{W2}$

So

$1 - B/C_{W2} \leq p \leq B/C_{W2}$

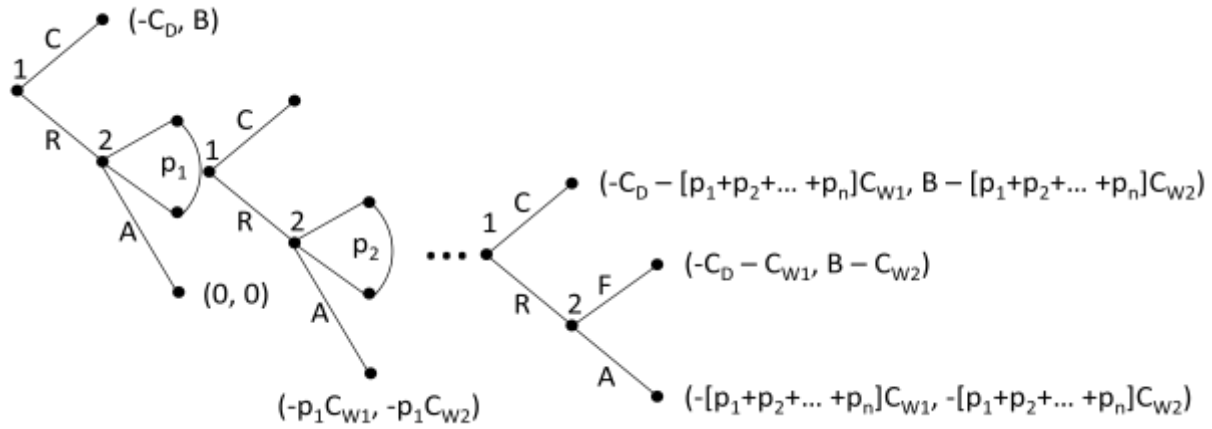
The deterrer capitulates at each node because the choice is always between a) losing after enduring the costs of war or b) capitulating and avoiding the costs of war. This outcome is therefore possible so long as $C_{W2} \leq 2B$. When $C_{W2} < 2B$, a range of p choices yield the same outcome. Because these are off the equilibrium path, it makes no difference which is chosen. Wherever any p is possible in equilibrium, p=.5 suffices.

Equilibria Outcomes:

$(-C_D, B)$ if $C_{W2} \leq 2B$ (Deterrer Capitulates)

$(0, 0)$ if $C_{W2} > 2B$ (Challenger Abandons the Challenge)

Figure 3: N Decision-Points



Solution

Challenger chooses p_1 if

$B \geq p_1C_{W2}$

Challenger chooses p_2 if

$B \geq p_2C_{W2}$

…

Challenger chooses p_n if

$B \geq p_nC_{W2}$

Challenger chooses F if

$B \geq (1 - [p_1+p_2+\ldots+p_n])C_{W2}$

Equilibria Outcomes:
$(-C_D, B)$ if $C_{W2} \leq (n+1)B$ (Deterrer Capitulates)
$(0, 0)$ if $C_{W2} > (n+1)B$ (Challenger Abandons the Challenge)
If $n = \infty$, deterrence never works

If many decision points exist after war begins, deterrence by punishment collapses completely.

Conclusions

This paper has proven that deterrence by punishment never works. It is time for the field to discard outmoded theories and concepts of deterrence.

No, not really. This paper is meant to be provocative, not to be taken literally. To preclude misunderstanding, I reject the conclusion that deterrence never works, even under the scope conditions necessary for this model to apply. Nonetheless, the model provides a surprisingly convincing proof that deterrence never works among perfectly rational actors if the would-be challenger has the ultimate ability to prevail militarily, regardless of the costs of doing so. In fact, this logic extends to any deterrence situation in which the challenger has the ability to gain by unilateral action in violation of the deterrent demand.[3] For example, states such as Iran or North Korea might use this strategy of exploiting sunk costs to defeat deterrent demands that they not build nuclear weapons.

I remain uncertain as to what to conclude from the model. Perhaps the field needs to revisit its assumptions about rational deterrence. Or perhaps the model says far more about the pitfalls of misusing backwards induction than it does about the effectiveness of deterrence. The errant assumption here may be perfect rationality taken to its (il)logical extreme. Another possibility is that deterrence by punishment is less important than has been assumed, with deterrence by denial corresponding more important. Even so, and even it is more wrong that right, the model would seem to capture a dynamic with the potential to make deterrence harder. The model also suggests a new logic for salami tactics as a strategy premised on exploiting sunk costs that deserves further inquiry.[4]

[3] In that sense, the conclusions apply to deterrence by punishment but not deterrence by denial. Glenn H. Snyder, *Deterrence and Defense* (Princeton: Princeton University Press, 1961), pp. 12-16.
[4] On salami tactics, see Schelling, *Arms and Influence*. For perfectly informed and rational actors, these incremental gains would always take place off the equilibrium path, but that does not diminish its potential effect on the outcome.