2011

Endogenizing the Sticks and Carrots: Modeling Possible Perverse Effects of Counterterrorism Measures

Vicki M. Bier
University of Wisconsin–Madison, bier@engr.wisc.edu

Kjell Hausken
University of Stavanger, kjell.hausken@uis.no

Ann Oper Res (2011) 186:39–59
DOI 10.1007/s10479-010-0833-8

Endogenizing the sticks and carrots: modeling possible perverse effects of counterterrorism measures

Vicki M. Bier · Kjell Hausken

Published online: 22 January 2011

Abstract We present a novel model capable of distinguishing between the effects of negative incentives ("sticks") and positive incentives ("carrots") for influencing the behavior of intelligent and adaptable adversaries. Utilities are developed for the defender and the terrorist. The defender is assumed to have a unit cost of defense, and unit costs of providing negative and positive incentives. The terrorist likewise has a unit cost of attack, which may either increase or decrease if the defender provides negative incentives, and enjoys a unit benefit of positive incentives. We show that the potential for perverse effects of counterterrorism (e.g., the emergence of hatred) can cause defenders to rely on positive incentives and decrease their reliance on negative incentives at equilibrium, with use of negative incentives completely eliminated in situations where these would be moderately effective when applied. With low potential for perverse effects of counterterrorism, the defender should rely on effective negative incentives.

Keywords Terrorism · Terror capacity · Threat · Conflict · Dynamics · Discounting · Contest success function

1 Introduction and background

In this paper, we develop a novel model for determining whether positive or negative sanctions are most likely to be effective in any particular case. We hope that this model will provide a framework within which this difficult question can be explored more rigorously than has typically been the case. In this respect, we view the current status of the "global war on terrorism" as analogous to the early days of the Cold War. Just as policy makers

V.M. Bier
Department of Industrial and Systems Engineering, University of Wisconsin-Madison, 1513 University Avenue, Room 3234, Madison, WI 53706, USA
e-mail: bier@cae.wisc.edu

K. Hausken (✉)
Faculty of Social Sciences, University of Stavanger, 4036 Stavanger, Norway
e-mail: kjell.hausken@uis.no

in the U.S. were originally unsure whether threatening to retaliate against a nuclear strike would make us safer or less safe from nuclear war, before work such as by Schelling (1960, 1978) and Aumann and Maschler (1995) forged a consensus that mutually assured destruction was in fact a stable equilibrium strategy, so too policy makers (and academics) today are divided over questions such as whether the war in Iraq is making the U.S. safer or less safe from Islamic terrorism, with no rigorous way to arrive at answers.[1] See Ganor (2005) for a discussion of some of the pros and cons of various policies.

1.1 Relevant academic literature

There is unfortunately scant relevant literature on this type of question. Bandyopadhyay and Sandler (2011) consider how preemption and defense interact with each other. For example, they note that high-cost defenders are likely to rely more heavily on preemption, while too little preemption "may exacerbate the [problem of] excessive defense... by making for an even more insecure environment." Elsewhere, however, Sandler et al. (2008) note that "offensive actions against terrorists and their supporters," while possibly effective at diminishing terrorism, do not seem to be cost effective (given their high expense relative to the current magnitude of the terrorism problem).

Bier et al. (2007) consider the option of making attacks more costly for terrorists, but do so in a way that treats carrots and sticks (positive incentives for cooperation, versus retaliation) symmetrically—an assumption which is almost certain to be violated in practice. In particular, Frey (2004; see also Frey and Luechinger 2003) points out some of the key differences between deterrence of terrorism through the threat of retaliation and other alternative strategies for protection, such as making oneself less vulnerable, or providing incentives for cooperation.

1.2 Positive incentives—general concepts and historical examples

Positive incentives ("carrots") could for example consist of providing goods, services, or valuable opportunities to groups that refrain from terrorism—making non-terrorist activity more attractive, as discussed by Frey (2004). Examples of positive incentives can include: direct monetary transfers; economic-development assistance; provision of goods, services, and opportunities; and removal of taxes or customs duties. Negative incentives ("sticks") are also intended to increase the cost of terrorism to potential terrorists, by such means as imposing trade restrictions, freezing the terrorist's assets, restricting how the terrorist operates, and retaliating militarily. However, we consider the possibility that such negative incentives could perversely alter the terrorist's intrinsic motivation to attack—e.g., through generating greater levels of hatred in the terrorists,[2] potentially leading them to redouble their efforts.[3]

[1] We owe this insight to remarks made by Richard Zeckhauser at a workshop on interdependent security in 2006, http://opim.wharton.upenn.edu/risk/IDS/publications.html.

[2] Regarding the emergence of hatred, Glaeser (2005) analyzes how politically motivated hatred of an "out group"—e.g., racism, anti-Semitism, or anti-Americanism (in the Islamic world)—can develop, when the out group has politically salient characteristics that distinguish it from the majority of the population (e.g., richer or poorer, disproportionately right-wing or left-wing, etc.). For example, he notes that "anti-Semitic hatred never became a significant political force" in countries where Jews were not primarily concentrated on one side of the political aisle. Thus, Glaeser explains how the emergence of hatred may depend on an out group's income level, predominant ideology, etc., but does not explicitly model the competition between a defender and a terrorist or other attacker.

[3] Positive and negative incentives may be interpreted as income effects. Enders and Sandler (2004) refer to "freezing terrorist's assets," which "reduces their 'war chest' and their overall ability to conduct a campaign of

In this paper, we do not explicitly discuss the issue of negotiating with terrorists.[4] However, we do address a closely related issue—namely, the conditions under which defenders would be willing to offer their adversaries positive incentives to refrain from attacking, and when potential attackers would be willing to accept such incentives. While controversial, this is not without precedent. For example, there is a well-established tradition of computer hackers being recruited as security experts. Presumably, the financial rewards and prestige of a successful career in computer security are sufficient to deter many former hackers from returning to their old ways (especially since their careers could be expected to continue only if they refrained from hacking), but the benefits of such collaborations also accrue to their employers. In accordance with Pruitt (2006), we anticipate that our approach would apply more readily when the demands of the terrorists are ones that can feasibly be met by the defenders (i.e., are not too high, or too inconsistent with the values of the defenders). If terrorist goals are relatively achievable in that sense (such as regional independence, or better treatment of an outcast minority), there will obviously be more opportunity to find accommodations that are mutually acceptable to both terrorists and defenders, since the level of positive incentives required to dissuade terrorists from future attacks could well be worthwhile in order to protect defenders from terrorist threats.

One argument against positive incentives is that such incentives may encourage attackers to continue fighting. In this paper, we assume that positive incentives are conditional, and hence provided only if the terrorist decides not to launch an attack. Of course, there is a risk that the defender may offer positive incentives and find that the terrorist still attacks. In this regard, we make an analogy with the idea of credible threats. In a complete game-theoretic analysis, a repeated game model would generally be needed to determine whether a particular threat is credible. For example, the defender may threaten massive retaliation if an attack is launched after the attacker has accepted positive incentives, but a multi-stage repeated game model would be required to determine when it would actually be in the defender's long-term interest to make good on that threat. However, in the more limited analysis conducted here, we consider a single-stage model, which could be viewed as a sub-game of a larger iterated game. Thus, the fact that the defender must have a credible threat of retaliation or other enforcement mechanism in order for our results to hold limits the applicability of our results—e.g., to situations in which retaliation is not too costly, etc.

terror"—for example, by freezing their bank accounts. Lakdawalla and Zanjani (2005) show that "protection reduces the payoff to terrorism"; they argue that "Deterrence [due to income reduction] takes place insofar as private self-protection raises [the level of non-violent activities] and lowers the total amount of violent terror investments." Finally, Hausken (2006) proposes a model in which defensive investment not only helps defend the defender's asset, but also reduces the terrorist's resources, so that the terrorist's attack effort becomes smaller.

[4]Conventional wisdom recommends not negotiating with terrorists or other similar adversaries (such as rogue states). However, such negotiations have of course occurred. In particular, Spector (1998, 2003) discusses negotiations between Israel and the Palestine Liberation Organization, between the U.S. and Haiti, between the U.S. and North Korea, and between Great Britain and Sinn Fein. In particular, Spector (2003) argues that "Despite the risks inherent in negotiating with terrorists, the risks of following a no-negotiation policy are likely to be more deadly. States need to assess terrorist interests and intentions to find if there are reasonable entry points for negotiation and take advantage of these to transform the conflict." Similarly, Pruitt (2006) considers both the peace process in Northern Ireland and negotiations with Islamic terrorist groups. He suggests that the success of negotiations depends on flexible attitudes on the part of both parties, and that "There are many arguments against negotiating with terrorists, but most of them do not apply to secret, backchannel talks, which are usually the method of choice in first approaching these groups." He also observes that "Negotiation with non-ideological ethno-nationalist terrorists is more common and more successful than with other kinds of terrorists."

Hoffman (2001), interviewing an Arafat loyalist (a former terrorist, a senior commander of al-Fatah, and a Palestine brigadier general) reports two examples of positive incentives. Although they were provided by the leadership of the terrorist organizations to discipline their organizations, rather than by the potential victims of terrorist attacks, the examples could be modified so that the potential victims could provide the positive incentives. First, after Hashemite King Hussein of Jordan sought to restore his monarchy's rule by quashing the autonomy of Palestinian organizations, killing tens of thousands of people (mostly Palestinians) from September 1970 to July 1971, Yasir Arafat formed the Black September Organization, the most elite unit of the Palestine Liberation Organization (PLO), consisting of dedicated, ruthless, loyal, and highly skilled warriors. Their first two operations were the November 1971 assassination of Jordan's Prime Minister Wasfi al-Tal (where one of the assassins knelt and lapped with his tongue the blood flowing across the marble floor), and the September 1972 seizure of Israeli athletes at the Munich Olympic Games (exemplifying terrorism's ability to transform a cause from obscurity to renown). Two years thereafter, Arafat was invited to address the General Assembly of the United Nations (UN), and thereafter the PLO was granted special UN observer status. Having obtained international recognition, Arafat wanted to "turn Black September off" (Hoffman 2001). The PLO leadership used positive incentives to recruit approximately 100 attractive young Palestinian women to Beirut. As Hoffman reports, "the hundred or so Black Septemberists were told that if they married these women, they would be paid $3,000; given an apartment in Beirut with a gas stove, a refrigerator, and a television; and employed by the PLO in some nonviolent capacity. Any of these couples that had a baby within a year would be rewarded with an additional $5,000." Thereafter, the PLO provided periodic tests of these individuals' willingness to return to terrorism, but none strayed, and Black September had been effectively dismantled.

Second, the authorities in Northern Ireland pursued a similar strategy before the 2001 cease-fire: "Hard-core IRA and Loyalist terrorists [mostly in their thirties] serving long prison sentences were often given brief furloughs during holiday periods" (Hoffman 2001). Combined with a variety of factors in prison conditions, and the possibility of early release, the objective was to allow these to develop family ties, and "wean these men from terrorism." According to Hoffman, "The program was so successful that the option could be offered to only a limited number of prisoners, lest the terrorist organizations, fearing the loss of too many senior veterans and commanders, forbid their members to participate in the program." The lesson to be learned, Hoffman (2001) argues, is that "creative thinking can sometimes achieve unimaginable ends." "Rather than concentrating on eliminating organizations, as we mostly do in our approach to countering terrorism, we should perhaps focus at least some of our attention on weaning individuals from violence."

As an amusing example of positive incentives, consider the following statement by Butch (Paul Newman) in the classic 1969 movie, "Butch Cassidy and the Sundance Kid" (Dirks 1969). Butch criticizes E. H. Harriman's expensive efforts to eliminate their menace once and for all as follows: "A set-up like that costs more than we ever took.... That crazy Harriman. That's bad business. How long do you think I'd stay in operation if every time I pulled a job, it cost me money? If he'd just pay me what he's spending to make me stop robbin' him, I'd stop robbin' him."

1.3 Negative incentives—general concepts and historical examples

Negative incentives are intended by the defender to make attacks more costly for the terrorist, in an attempt to induce the terrorist to refrain from attacking. Examples of negative incentives might include: military retaliation; termination of trade relations with countries

that engage in or sponsor terrorist activities (either directly or indirectly—e.g., by providing safe havens for terrorists); imposition of (selective) trade restrictions; freezing of terrorist assets; limits on the movement of personnel, goods, and services by terrorist organizations; and restrictions on the manner in which terrorist groups can operate.

We assume that negative incentives may be experienced either negatively or positively by terrorists. In other words, the threat of negative incentives may have either a deterrent impact, or a variety of perverse effects such as the emergence of hatred or the desire for vengeance among terrorists and their supporters, the rise of hostile groups, loss of international prestige, political instability in allied nations, etc. Perverse effects of counterterrorism may alter terrorists' intrinsic motivation, increasing their willingness to attack. Terrorists may get the feeling that they are fighting with their backs against the wall (i.e., in "death ground"; Tzu 320 Before Christ). This may inspire a redoubling of the terrorists' effort.

For example, the imposition of negative incentives could increase recruitment to terrorism, or increase political support for the methods and objectives of the terrorists (which in turn could generate additional sponsors and economic support for terrorism). Lutfi (2001) thus argues that "containing the Xinjiang Islamist threat is likely to backfire, much in the same way containment of the Soviet Union in Afghanistan produced the ongoing blowback of terrorism." This is modeled as a decrease in the terrorists' unit cost of attack, since some of the terrorists' activity may "come for free" as a result of the increased intrinsic motivation and/or capability to attack. For example, it seems reasonable to assume that increasing the availability of recruits, funds, or safe havens would enable a terrorist group to function more efficiently; instead of using poorly trained personnel and low-quality equipment, a terrorist group may be able to upgrade its personnel and equipment to operate more capably, with a lower unit cost of attack. Thus, in this paper, we assume that ample availability of resources enables a terrorist to operate with a lower unit cost of attack (by analogy with the idea that greater supply leads to lower price). In a constrained variant of our model, it might also be possible to represent the perverse effects of negative incentives as an increase in the budget constraint; while we have not analyzed this version of the model, we anticipate that it would likely lead to generally similar results.

Determining when negative incentives are likely to have a deterrent impact on terrorists, and when they are likely to arouse greater antagonism, will obviously be extremely difficult in practice. In fact, even terrorists may not know when negative incentives are first imposed whether they will react with submission or vengeance. However, work like that of Glaeser (2005) and Sandler et al. (2008) could eventually help to quantify the impact of negative incentives (or at least help in determining whether they are likely to be experienced as negative or positive).

We assume that the terrorist has a unit cost of attack which can increase or decrease through the imposition of negative incentives. For a historic example where it is reasonable to assume that negative incentives cause the terrorist's unit cost of attack to increase, consider U.S. relations with Libya. This relationship began to deteriorate when Gaddafi seized power in 1969. The U.S. withdrew its ambassador in 1972, and Libya was designated as a "state sponsor of terrorism" in 1979. The U.S. became increasingly frustrated over the lack of success of its covert attempts to topple Libyan president Gaddafi. Two weeks after the Berlin discotheque bombings, the U.S. sent 30 bombers to strike Tripoli and Benghazi for 11 minutes in the early morning of April 14, 1986.[5] The objective was to kill Gaddafi—who

[5]See Burton (2008) for an account of the events, and a discussion of whether the attack was the culmination of a long-term US policy to change the Libyan regime, and/or a tactical retaliation to the Berlin discotheque bombings.

had been dubbed a "mad dog" by then President Reagan. The immediate outcomes were: survival of Gaddafi; hospitalization of his wife and eight of his children (some with serious injuries); and the death of his adopted daughter (aged 15 months). In retrospect, these negative incentives appear to have been successful at deterring future attacks (increasing the terrorist's unit cost of attack), since after the incident, Libya disappeared from media attention as a sponsor of terrorist attacks, while the U.S. and Britain encouraged and funded opposition groups within the country. Eventually, Gaddafi pledged his support for the "war against terrorism," and agreed to pay compensation to the victims of the 1988 Lockerbie bombing (for which a Libyan intelligence agent had been jailed). Formal U.S. relations with Libya were reestablished in 2006, and U.S. Secretary of State Rice visited Gaddafi on September 5, 2008—the first such visit since 1953.

Conversely, for historic examples where negative incentives can possibly cause the terrorist's unit cost of attack to decrease (inspiring perverse effects such as the emergency of hatred and/or redoubling of attack effort), consider the following. First, Euskadi Ta Askatasuna (a Basque separatist organization) has killed over 800 people since 1968, despite strong negative incentives from Spain, including the current incarceration of more than 700 ETA members.[6] Second, Al-Qaeda continues to experience recruitment of new members, despite substantial efforts to curtail its activities following the attacks of September 11, 2001. Third, considering China as the "defender" and Tibetan-rights activists as "attackers," Chinese opposition to Tibetan freedom appears to have strengthened the Tibetan Youth Congress (founded by descendents of the 14th Dalai Lama and other Tibetan aristocrats in exile); this organization, established in 1970, now has about 30,000 members, and up to 70 branches worldwide.[7] Fourth, Hezbollah (founded in 1982, and considered by many to be a terrorist organization) has built itself up over time, with lavish support from Syria and Iran, despite heavy opposition and negative incentives from countries such as Israel. More generally, the list of active terrorist organizations continues to be long, despite substantial negative incentives against terrorism over a period of decades.

1.4 Overview of the paper

Section 2 presents the model. Section 3 analyzes the model without positive incentives. Section 4 illustrates the results. Section 5 presents conditionality tests for the desirability of positive incentives. Section 6 concludes.

2 The model

A terrorist intends to destroy something of value, which we refer to as an asset. The attack may cause physical destruction, loss of lives, injuries, fear, etc.,[8] which we include in the value of the asset. The asset can thus have joint economic, human, and symbolic value. We consider a contest between a terrorist and a defender over this asset, which is owned by the defender. The defender chooses a level of defensive investment with which to protect the

[6] El Confidencial, January 7, 2008, http://www.elconfidencial.com/cache/2008/01/01/66_cifra_presos_ultima_decada_encarcelados.html.

[7] http://www.globalresearch.ca/index.php?context=va&aid=8691.

[8] See Hoffman (1998) regarding terrorist objectives.

asset, and the terrorist chooses a level of effort with which to attack the asset.[9] Additionally, the defender can provide positive and/or negative incentives to influence the terrorist's behavior, at some cost to itself.

Consider a defender with an asset that a terrorist seeks to destroy. The defender and the terrorist can each be an individual, a country, an organization, a firm, a group, etc. The defender values the asset as r, and the terrorist values the asset as R. Alternatively, the terrorist may assign the value R to successful completion of a terrorist operation against that asset. The defender chooses an effective level of defense t with which to defend its asset, and the corresponding defensive expenditure is f, where $\partial f/\partial t > 0$. The terrorist chooses an effective level of terrorist effort T with which to attack the asset, with corresponding attack expenditure F, where $\partial F/\partial T > 0$. We consider the simple cases $f = at$ and $F = AT$, where the expenditures can be in the form of capital and/or labor, and the parameters a and A are the unit costs of defense and attack. For simplicity, we assume risk-neutrality; however, this does not change the fundamental nature of our argument.

The contest between the defender and the terrorist for the asset takes the common ratio form (Skaperdas 1996; Tullock 1967). In particular, we consider the contest success function

$$h = \frac{t}{t+T} \tag{1}$$

which can be interpreted either as the probability that the defender retains its asset, or as the fraction of its asset that the defender retains. For simplicity, we use the probability terminology throughout this paper, but point out that either interpretation is equally valid. Equation (1) satisfies $\partial h/\partial t > 0$ and $\partial h/\partial T < 0$. That is, the defender benefits from its defense, and suffers from the terrorist's attack.

For the basic model, we formulate the defender and terrorist utilities u and U as

$$u = \frac{t}{t+T}r - at, \qquad U = \frac{T}{t+T}R - AT \tag{2}$$

Both actors incur costs from their expenditures, but also enjoy benefits—the defender by increasing the probability of keeping its asset, and the terrorist by increasing the probability of destroying the asset.

We now generalize this basic model by introducing two additional effects, labeled "carrots" and "sticks" (i.e., positive and negative incentives or sanctions, p and n, respectively), provided by the defender to affect the terrorist's behavior. The defender's costs of providing positive and negative incentives are given by cp and bn. For example, one can think of the expenditures cp and bn as consisting of capital and/or labor; however, incentives can also be provided by changing laws and regulations, engaging in diplomacy or lobbying, etc.

Positive incentives are assumed to be conditional, in the sense that they are provided only if the terrorist decides not to launch an attack (i.e., if $T = 0$). We interpret such positive

[9]The contest can take place anywhere, such as in the defender's backyard, on neutral ground, or in the terrorist's backyard. It may involve activities of a military, political, economic, or other nature, interpreted broadly. The reason for this broad interpretation is that the terms "attack" and "defense" can be understood as metaphors. As Hirshleifer (1995) puts it, "falling also into the category of interference struggles are political campaigns, rent-seeking maneuvers for licenses and monopoly privileges (Tullock 1967), commercial efforts to raise rivals' costs (Salop and Scheffman 1983), strikes and lockouts, and litigation—all being conflictual activities that need not involve actual violence." Attack and defense are thus best viewed as subcategories of attack-oriented and defensive competition. In this paper, we use the relatively narrow terms of "attack" and "defense," but these can be replaced with broader terms such as struggle or conflict as desired.

incentives p as being unequivocally favorable for the terrorist; their impact is modeled as a benefit of Cp if the terrorist doesn't launch an attack, $T = 0$, where C is the unit benefit of the positive incentives.

We model negative incentives n as altering the terrorist's unit cost of launching an attack from A to $A + Bn^k$, where B describes the magnitude and direction of the effect of negative incentives, and $k > 0$ is an additional parameter describing the increasing or diminishing returns to increased negative incentives. We hereafter refer to A as the basic unit cost for the terrorist. The parameter B is assumed to be exogenously given, and is obviously not straightforward to estimate (as discussed above). A defender might choose to implement negative incentives in the hope that B is positive (so that the negative incentives will effectively increase the unit cost of attacking), but may not be able to ensure that B is positive. The case $0 < k < 1$ implies a concave increase in the unit cost as a function of n; from the point of view of the defender, this means that negative incentives are subject to diminishing returns in reducing the terrorist's unit cost. Conversely, $k > 1$ means increasing returns to negative incentives.

A unit cost of $A + Bn^k$ means that the terrorist's total cost of launching an attack of magnitude T is $(A + Bn^k)T$, which implies that n and T have a multiplicative impact. Similar reasoning has been applied by Dalvi et al. (2004), in the absence of a contest success function. They argue that the costs of investments by one agent increase proportionally with the level of investments made by the adversary, because every additional unit of investment is now that much less "effective," due to the adversary's investment. Applied to our model when B is positive, this idea suggests that negative incentives n provided by the defender causes the terrorist's attack to be less effective. According to Dalvi et al., the agent's cost expenditure is proportional to both n and T. More generally, in our case, we allow the cost of an attack to change by an amount Bn^kT, relative to its initial value of AT.

For a historic example where it is reasonable to assume that the parameter B was positive, consider U.S. relations with Libya, as discussed in the introduction. Conversely, for examples of situations in which the parameter B was plausibly negative, consider the examples toward the end of the introduction. If the overall magnitude of the perverse effects (e.g., hatred) is greater than the direct effect of the negative incentives, then the effect of the negative incentives can be represented as causing an overall reduction in the terrorist's unit cost below A. In other words, the terrorist may be able to generate the same or higher level of attack effort T more cheaply, more cost efficiently, with more support, with fewer obstacles, and potentially from within a safe haven (where the terrorist may enjoy economies of scale in the development of its operations). This means that whereas A is always positive, the parameter B can in principle be negative, causing a reduction to a unit cost below A for the terrorist.[10] Of course, if the defender knew that B was negative, it would choose $n = 0$ and avoid the use of negative incentives. Therefore, in the analysis below, we focus on what happens to the equilibrium defender and terrorist strategies as B approaches 0 from above.

To clarify the difference between the defensive effort t and the negative sanctions n, the defense t is directed toward increasing the probability $t/(t + T)$ of the defender keeping its asset, which the defender prefers to be as large as possible. In principle, this so-called "defense" can be either defensive or offensive in nature; the distinctive feature for the purposes of our model is that the defense t is matched against the terrorist's attack effort T in a contest to determine with what probability the defender retains its asset. Unlike the defenses t, the

[10]It seems reasonable to assume that the term $A + Bn^k$ will usually be positive in practice, even when B is negative. However, there is no reason in principle that this term cannot be negative; in such cases, the terrorist would essentially get paid to launch attacks, benefiting more when the attack effort T is high.

negative incentives n are not injected directly into the contest. That is, they are not matched against the attack effort T, and (if T is held constant) do not help the defender increase its probability of keeping its asset. Instead, negative incentives are directed at altering the terrorist's unit cost of launching an attack (and thereby affecting its choice of T).

Summarizing the discussion above, we generalize the defender and terrorist utilities (u and U, respectively) from (2) as follows:

$$\begin{aligned} u &= \begin{cases} \frac{t}{t+T} r - at - bn - cp & \text{if } T = 0 \\ \frac{t}{t+T} r - at - bn & \text{if } T > 0 \end{cases} \\ U &= \begin{cases} \frac{T}{t+T} R - (A + Bn^k)T + Cp & \text{if } T = 0 \\ \frac{T}{t+T} R - (A + Bn^k)T & \text{if } T > 0 \end{cases} \end{aligned} \tag{3}$$

where $a > 0$, $b > 0$, $c \geq 0$, $C \geq 0$, and B can be either positive or negative. This formulation means that the defender has three strategic choice variables (t, n, and p), while the terrorist has one strategic choice variable (T). We assume that all parameters are common knowledge. Of course, many of the parameters, most importantly B, are challenging to estimate, to say the least. Therefore, we put forward this model not as a definitive way to determine what defenders should do, but rather as a building block for future models in which the defender may have probability distributions over some of the key model parameters.

In the following sections, we analyze a two-period game and solve for sub-game perfect equilibrium. In the first period, the defender chooses the defense t and negative incentives n, with the expectation that the terrorist chooses an equilibrium value of T in the second period, depending on the values of t and n chosen in the first period. In the second period, the terrorist chooses T. Finally, a conditionality test is applied to determine the level of positive incentives (if any) chosen by the defender. Letting the defender move first is frequently realistic in practice, since the defender may for example build up infrastructure and defensive systems over time. The terrorist organization may either not yet exist during this initial phase of the game, or may take a wait-and-see approach; it eventually takes the infrastructure as given when choosing its optimal strategy in the second period. The game is solved by backward induction.

3 Analyzing the model without positive incentives ($c = \infty$ or $C = 0$)

We begin by analyzing the model in the absence of positive incentives. Assuming sub-game perfect equilibrium, we first solve for the equilibrium value of the attack effort T in the second period, conditional on t and n in the first period. We thereafter find the optimal t and n in the first period, taking into account that the terrorist's choice of T in the second period must be in equilibrium.

The terrorist's first-order condition in the second period is determined by differentiating U in (3) with respect to T, and setting the result equal to zero:

$$\frac{\partial U}{\partial T} = \frac{tR}{(t+T)^2} - (A + Bn^k) = 0 \quad \Rightarrow \quad T = \sqrt{\frac{tR}{A + Bn^k}} - t \tag{4}$$

Inserting the result into (3) and simplifying gives the defender's first-period utility as

$$u = \frac{r}{\sqrt{R}} \sqrt{t} \sqrt{A + Bn^k} - at - bn \tag{5}$$

Equation (5) illustrates that the defender has the option to provide negative incentives in the first period. Differentiating u in (5) with respect to n when $B > 0$ gives $n > 0$; by contrast, differentiating u in (3) with respect to n when $B > 0$ (which would correspond to a simultaneous-move game) would give $n = 0$. As noted above, when $B < 0$, the defender maximizes u by setting $n = 0$, which increases the first term in (5), and eliminates the third term (thus also eliminating the potential for the terrorist to benefit from a unit attack cost less than A). Differentiating (5) with respect to t and n gives the following first-order conditions:

$$\frac{\partial u}{\partial t} = \frac{r\sqrt{A + Bn^k}}{2\sqrt{R}\sqrt{t}} - a = 0, \qquad \frac{\partial u}{\partial n} = \frac{Bkn^{k-1}r\sqrt{t}}{2\sqrt{R}\sqrt{A + Bn^k}} - b = 0 \tag{6}$$

Solving for the defender's equilibrium strategy gives

$$t = \begin{cases} (A + B(\frac{4bRa}{Bkr^2})^{\frac{k}{k-1}})\frac{r^2}{4Ra^2} & \text{when } B > 0, \\ \frac{Ar^2}{4Ra^2} & \text{when } B \le 0, \end{cases} \qquad n = \begin{cases} (\frac{4bRa}{Bkr^2})^{\frac{1}{k-1}} & \text{when } B > 0, \\ 0 & \text{when } B \le 0 \end{cases} \tag{7}$$

Inserting the results into (4) then gives

$$T = \begin{cases} (1 - \frac{Ar}{2Ra} - \frac{Br}{2Ra}(\frac{4bRa}{Bkr^2})^{\frac{k}{k-1}})\frac{r}{2a} & \text{when } B > 0 \\ (1 - \frac{Ar}{2Ra})\frac{r}{2a} & \text{when } B \le 0 \end{cases} \tag{8}$$

Finally, inserting the defender and terrorist equilibrium strategies into the expressions for u and U in (3), confirmed by u in (5), gives

$$\begin{aligned} u &= \begin{cases} (A - B(k-1)(\frac{4bRa}{Bkr^2})^{\frac{k}{k-1}})\frac{r^2}{4Ra} & \text{when } B > 0 \\ \frac{Ar^2}{4Ra} & \text{when } B \le 0 \end{cases} \\ U &= \begin{cases} (1 - \frac{Ar}{2Ra} - \frac{Br}{2Ra}(\frac{4bRa}{Bkr^2})^{\frac{k}{k-1}})^2 R & \text{when } B > 0 \\ (1 - \frac{Ar}{2Ra})^2 R & \text{when } B \le 0 \end{cases} \end{aligned} \tag{9}$$

Note, of course, that (7) provides only an interior solution for t and n when $B > 0$. It is straightforward to see from (5) that choosing $t = 0$ and $n > 0$ would give $u < 0$, but the defender would still need to determine whether the corner solution corresponding to $t > 0$ and $n = 0$ yields higher utility than the interior solution in (7). Solving the first equation in (6) for $n = 0$, and inserting the result into (3) and (4), gives

$$t = \frac{Ar^2}{4Ra^2}, \qquad n = 0, \qquad T = \left(1 - \frac{Ar}{2Ra}\right)\frac{r}{2a}, \qquad u = \frac{Ar^2}{4Ra}, \qquad U = \left(1 - \frac{Ar}{2Ra}\right)^2 R \tag{10}$$

When $1 \le Ar/(2Ra)$, the terrorist would choose $T = 0$, provided that the defender chooses t large enough to ensure that $T = 0$ in the second period. In particular, inserting $T = n = 0$ into (4) gives $t = R/A$. Generalizing (10) to allow for that case gives

$$\begin{aligned} t &= \begin{cases} \frac{Ar^2}{4Ra^2} & \text{when } 1 > \frac{Ar}{2Ra}, \\ \frac{R}{A} & \text{when } 1 \le \frac{Ar}{2Ra}, \end{cases} \quad n = 0, \quad T = \begin{cases} (1 - \frac{Ar}{2Ra})\frac{r}{2a} & \text{when } 1 > \frac{Ar}{2Ra}, \\ 0 & \text{when } 1 \le \frac{Ar}{2Ra}, \end{cases} \\ u &= \begin{cases} \frac{Ar^2}{4Ra} & \text{when } 1 > \frac{Ar}{2Ra}, \\ r - \frac{Ra}{A} & \text{when } 1 \le \frac{Ar}{2Ra}, \end{cases} \quad U = \begin{cases} (1 - \frac{Ar}{2Ra})^2 R & \text{when } 1 > \frac{Ar}{2Ra} \\ 0 & \text{when } 1 \le \frac{Ar}{2Ra} \end{cases} \end{aligned} \tag{11}$$

When $B > 0$ and $1 > Ar/(2Ra)$, then the defender would prefer to choose the solution from (9) when $0 < k < 1$, and the solution from (11) when $k \geq 1$. We now formulate this result as a proposition below:

Proposition 1 *Assume $B > 0$ and $1 > Ar/(2Ra)$. The defender provides negative incentives (i.e., chooses $n > 0$) when $0 < k < 1$, and does not provide negative incentives (i.e., $n = 0$) when $k \geq 1$.*

Proof The expression for u in (11) is larger than or equal to the expression for u in (9) when

$$\frac{Ar^2}{4Ra} \geq \left(A - B(k-1)\left(\frac{4bRa}{Bkr^2}\right)^{\frac{k}{k-1}}\right)\frac{r^2}{4Ra} \quad \Rightarrow \quad k \geq 1 \tag{12}$$

□

Proposition 2 *When $B > 0$, $1 > Ar/(2Ra)$, and $0 < k < 1$, then we have $\partial t/\partial B > 0, \partial n/\partial B > 0, \partial T/\partial B < 0, \partial u/\partial B > 0$, and $\partial U/\partial B < 0$.*

Proof The result follows from differentiating (7), (8), and (9), shown in the Appendix. □

The Appendix shows how three choice variables t, n and T, and the two outcome variables u and U, vary with k. The role of the fourth choice variable, p, is considered in Sect. 5. When $k \geq 1$, the expressions for T and n are increasing in k, while t and u are decreasing in k. Thus, increasing returns to negative incentives for the defender cause the terrorist to increase its attack effort T (because if the terrorist were to respond to increased unit cost $A + Bn^k$ for k large with low T, it would not succeed in the contest, but would still incur a high cost). Foreseeing that the terrorist will respond with a large T, the defender therefore chooses not to use negative incentives when k is large. In other words, values of $k \geq 1$ are a double edged sword that are not actually to the advantage of the defender. Since setting $k \geq 1$ causes $n = 0$, we confine our attention to $0 < k < 1$ in this paper.

When $B > 0$, then t, n, and U are always non-negative; solving (8) and (9) for T and u can result in negative values, but under opposite circumstances. Solving (8) for T will give a negative value when the terrorist's unit cost of effort A is large, which makes terrorism too costly; in this case, we assume that the terrorist would choose the corner solution $T = 0$. Conversely, solving (9) for u will give a negative value when A is small, in which case the equilibrium terror attack would be so overwhelming that the defender would not gain enough from the contest to justify the costs of t and n; in this case, the defender would set $t = n = 0$, giving a corner solution. In other words, terrorism is deterred if A is large, and the defender gives up on defending itself if A is small.

Proposition 3 *Let*

$$\begin{aligned} A_D &= B(k-1)\left(\frac{4bRa}{Bkr^2}\right)^{\frac{k}{k-1}}, \\ A_T &= \frac{2Ra}{r} - B\left(\frac{4bRa}{Bkr^2}\right)^{\frac{k}{k-1}} = \frac{2Ra}{r}\left(1 - \left(\frac{2Ra}{Br}\right)^{\frac{1}{k-1}}\left(\frac{2b}{kr}\right)^{\frac{k}{k-1}}\right) \end{aligned} \tag{13}$$

When $A > A_T$, the terrorist chooses $T = 0$, and the defender keeps its entire asset; otherwise, the terrorist attacks, and obtains part of the value of the asset. When $A < A_D$, the

defender gives up, chooses $t = n = 0$, *and loses its entire asset*; *the defender earns positive utility, and retains part of its asset, when* $A_D < A < A_T$.

Proof The results follow from requiring $T > 0$ in (8), and $u > 0$ in (9), respectively. □

Since we confine attention to the case where $0 < k < 1$, (13) gives $A_D < 0$, which means that the defender never withdraws from the contest. When $A_T < 0$, the disadvantaged terrorist always withdraws, since $A > 0$. The case $A_T > 0$ gives a range $0 < A < A_T$ where the agents share the asset, and a range $A > A_T$ where the terrorist is deterred from attacking.

Both A and B are ingredients of the terrorist's unit cost of attack, but unlike A, the effect of B is multiplied by n^k, and can even be negative. Proposition 3 can thus be reformulated as follows:

Proposition 4 *Let*

$$B_D = \left(\frac{4bRa}{kr^2}\right)^k \left(\frac{k-1}{A}\right)^{k-1}, \qquad B_T = \frac{2Ra}{r}\left(\frac{2b}{kr}\right)^k \left(1 - \frac{Ar}{2Ra}\right)^{1-k} \tag{14}$$

When $B > B_T$, *the terrorist chooses* $T = 0$ *and the defender keeps its entire asset*; *by contrast, the terrorist attacks* (*choosing* $T > 0$) *when* $B < B_T$. *When* $B < B_D$, *the defender gives up its entire asset, choosing* $t = n = 0$; *the defender earns positive utility when* $B_D < B < B_T$.

Proof The results follow from requiring $u > 0$ in (9), and $T > 0$ in (8), respectively. □

Analogously, since we are assuming $0 < k < 1$, (14) gives $B_D < 0$. The two cases $B < B_T < 0$ and $B_T < B < 0$ are different from $A_T < 0$, since B can be negative. However, the defender responds to negative B with $n = 0$, in which case (10) applies and the terrorist withdraws, since having $B_T < 0$ in (14) causes $1 > Ar/(2Ra)$. As with $A_T > 0$, the case $B_T > 0$ gives a range $0 < B < B_T$ where the agents share the asset, and a range $B > B_T$ where the disadvantaged terrorist is deterred. (As we proceed, we'll focus on how the agents share the asset when B decreases toward 0, which illustrates the emergence of perverse effects such as hatred.)

Finally, when $A > A_T$ or $B > B_T$ (so that the terrorist is deterred, and chooses $T = 0$), the defender must maintain $t = R/(A + Bn^k)$ to ensure the equilibrium solution of $T = 0$ in (4). The defender must solve $t = R/(A + Bn^k)$ together with the first equation in (6), then solve $t = R/(A + Bn^k)$ together with the second equation in (6), and choose the solution which gives the highest utility. (These three equations cannot be solved simultaneously, since the system is over-determined, with only two unknowns, t and n; the defender cannot solve for both t and n while maintaining $t = R/(A + Bn^k)$.) Solving $t = R/(A + Bn^k)$ together with the first equation in (6) gives

$$t = \frac{r}{2a}, \qquad n = \left(\frac{1}{B}\left(\frac{2Ra}{r} - A\right)\right)^{\frac{1}{k}}, \qquad u = \frac{r}{2} - b\left(\frac{1}{B}\left(\frac{2Ra}{r} - A\right)\right)^{\frac{1}{k}}, \qquad T = U = 0 \tag{15}$$

Observe that t and n here depend on a, but not on b. Equation (15) gives a value of t that accounts for a, and is optimal when b is low but not when b is high, since sufficiently large values of b will eventually cause negative utility u. When b is large, the defender solves

$t = R/(A + Bn^k)$ together with the second equation in (6). Those two equations are not analytically solvable for general k, but when $k = 1/2$ the solution is given by

$$t = \frac{2R}{\sqrt{A^2 + B^2 r/b} + A}, \qquad n = \left(\frac{\sqrt{A^2 + B^2 r/b} - A}{2B}\right)^2,$$
$$u = r - at - bn, \qquad T = U = 0, \qquad k = \frac{1}{2} \tag{16}$$

Now, t and n depend on b, but not on a. Equation (16) gives a value of n that accounts for b and is optimal when a is low, but not when a is high, since sufficiently large values of a will eventually cause negative utility u. The defender chooses to solve either (15) or (16), depending on which one gives the higher utility.

4 Illustrating the results of the model without positive incentives ($c = \infty$ or $C = 0$)

Figures 1 and 2 plot the five variables characterizing the equilibrium solution to our model (t, n, T, u, and U) as functions of the terrorist's unit attack cost A, and the response to negative incentives B. The legend for each figure specifies which variables are plotted, so there are no labels on the vertical axis. All figures were produced using baseline parameter values $a = b = A = B = 1$, $r = R = 4$, $k = 1/2$. With these parameter values, (13) gives $A_T = 3/2$, (14) gives $B_T = \sqrt{2}$, and the defender prefers (16) over (15).

Figure 1 shows how the characteristics of the equilibrium solution change with the unit attack cost A. When $A > 1.5 = A_T$, then in accordance with Proposition 3, the terrorist withdraws, and the defender never gives up on protecting its asset, since A_D is negative. In accordance with (7)–(9), characterizing the interior solution that exists when $A < 1.5$, increases in the unit cost of attack A yield increases in the defender's investment t and utility u, and decreases in the terrorist's effort T and utility U. The level of negative incentives n is constant in A, since changing the unit cost of attack A does not affect the terrorist's response to negative incentives. In accordance with (15) when $A > A_T = 1.5$, the defender chooses t and n optimally while deterring the terrorist.

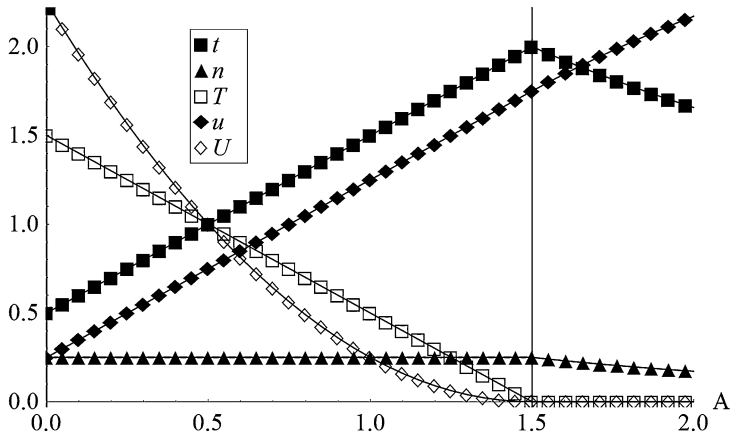

Fig. 1 Behavior of the equilibrium solution as a function of the unit attack cost A

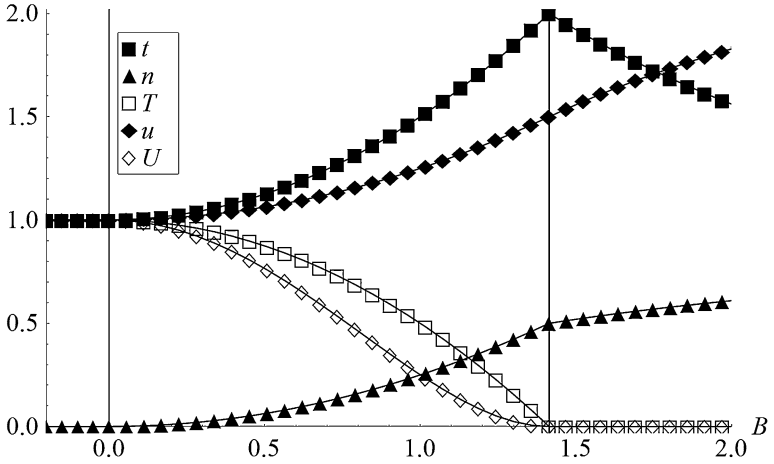

Fig. 2 Behavior of the equilibrium solution as a function of the incentive parameter B

Figure 2 characterizes the behavior of the equilibrium solution as a function of B when $k = 0.5$. This figure shows that negative sanctions can be an effective component of an overall defensive strategy (increasing the effectiveness of defensive investments t) under some circumstances. In fact, when B is sufficiently large, $B > \sqrt{2} = B_T$ (in accordance with Proposition 4), the threat of negative sanctions deters attacks altogether. However, from Proposition 2, as B decreases, the defender decreases t and n, and suffers lower utility, while the terrorist increases T and enjoys higher utility. According to (7), the defender decreases the negative incentives n gradually toward zero as B decreases toward zero; i.e., as negative sanctions tend to produce more perverse effects (e.g., hatred), the defender gradually reduces the use of negative incentives. However, some non-zero level of negative incentives remains desirable as long as their net effect is in the desired direction (i.e., as long as $B > 0$). When $B \leq 0$, the defender naturally sets $n = 0$; at this point, the defender and terrorist levels of investment and utility no longer depend on the specific numerical value of B. Numerical analysis finds no dramatic differences as k varies above or below 0.5.

5 Conditionality tests for the desirability of positive incentives

We now introduce the possibility of using positive incentives to reward desired behavior. We begin with a proposition.

Proposition 5 *There exists an equilibrium with positive incentives, characterized by sufficiently low c and sufficiently high C, which eliminates the equilibrium without positive incentives.*

Proof When $c = \infty$, the defender does not provide positive incentives, since they would be infinitely costly; likewise, when $C = 0$, the terrorist cannot benefit from positive incentives, making them of no value in influencing terrorist behavior. Conversely, when $c = 0$, the defender can provide positive incentives at no cost; when $C = \infty$, the terrorist benefits infinitely from positive incentives. Assume that the extent of positive incentives p is arbitrarily small but positive. Then if $T = 0$, the terrorist's utility will be arbitrarily large when

C is arbitrarily large. When c is also arbitrarily small but positive, then the level of negative sanctions is set to $n = 0$, and the defender chooses $t > 0$ arbitrarily small but positive to win the contest. In this case, the defender's utility, $u = r - at - bn - cp$, will be arbitrarily close to r. Since both the terrorist and the defender benefit from positive incentives for sufficiently small c and large C, the equilibrium without positive incentives no longer arises. □

In the previous section, we effectively set $c = \infty$ or $C = 0$ in order to eliminate any role for positive incentives. However, Proposition 5 states that a sufficiently low unit cost c of positive incentives for the defender, and/or a sufficiently high unit benefit C of positive incentives for the terrorist, can make positive incentives worthwhile. Observe in (3) how a large C decreases the impact of a low B. This can enable the defender to eliminate the perverse effects of counterterrorism (e.g., hatred). This means that with low c and high C, there can now be a new equilibrium with positive incentives. To address this situation, we assume that the defender evaluates both games (one with the option of positive incentives, and one without), and plays whichever one is better (since the defender is the first mover). The default game is the one without positive incentives. The game with positive incentives is such that the defender offers the terrorist the smallest level of positive incentives p in the first period required to achieve $T = 0$ in the second period, i.e., no terrorist attack. This obviously requires the establishment of monitoring and enforcement capabilities by the defender, so that the terrorist does not simply claim positive incentives in the first period and then still attack in the second period. In this paper, we assume that the defender chooses positive incentives when optimal, and avoids positive incentives when they can't be adequately enforced. If the attacker attacks after having received positive incentives, the defender loses. We do not analyze this case, but it can be expected that the defender will then not provide positive incentives to the terrorist in the future. Analogously, we assume that if the defender offers positive incentives, and the terrorist does not attack, then positive incentives are provided to the terrorist. If the defender does not provide the positive incentives it has offered despite the terrorist not attacking, the terrorist loses. We do not analyze this case, but it can be expected that the terrorist will then not refrain from attacking in the future, due to not trusting that the defender will actually provide the offered positive incentives.

We define t_{eq}, n_{eq}, and T_{eq} as the equilibrium values of t, n, and T without positive incentives, and u_{eq} and U_{eq} as the corresponding equilibrium values of the utilities u and U. After t_{eq}, n_{eq}, T_{eq} have been computed, conditionality tests on u_{eq} and U_{eq} are performed to determine whether non-zero positive incentives p could eliminate the attack effort T_{eq}. The positive incentives can be perceived as an outside option, against which both agents compare their utilities in the two-period game. Because the parameters of the game are common knowledge, both agents will know whether the conditionality tests are satisfied, and will therefore know whether the defender will be willing to reward the terrorist for refraining from attacks. This means that when the positive incentives are provided, the two-period game is not played. If the terrorist chooses $T = 0$ in response to positive incentives p, its utility in (3) is then given by $U = Cp$. Therefore, the terrorist is willing to choose $T = 0$ rather than T_{eq} if

$$p > \frac{1}{C}\left(\frac{T_{eq}}{t_{eq} + T_{eq}} R - (A + Bn_{eq}^k)T_{eq}\right) = p_{min} \tag{17}$$

where p_{min} is the level of positive incentives that makes the terrorist indifferent between launching no attack ($T = 0$) and choosing T_{eq}. (This is clearly the minimum effective level of positive incentives, since the terrorist would prefer to get more positive incentives if

possible.) Inserting (7) and (8) into (17) gives

$$p_{min} = \begin{cases} (1 - \frac{Ar}{2Ra} - \frac{Br}{2Ra}(\frac{4bRa}{Bkr^2})^{\frac{k}{k-1}})^2 \frac{R}{C} & \text{when } B > 0 \\ (1 - \frac{Ar}{2Ra})^2 \frac{R}{C} & \text{when } B \leq 0 \end{cases} \tag{18}$$

The next question to address is whether the defender would be willing to provide p_{min}. From (3), the defender prefers to provide positive incentives p to obtain $T = 0$ rather than T_{eq} when

$$\frac{t_{eq}}{t_{eq}+0} r - at_{eq} - bn_{eq} - cp > \frac{t_{eq}}{t_{eq}+T_{eq}} r - at_{eq} - bn_{eq} \quad \Rightarrow \quad p < \left(\frac{T_{eq}}{t_{eq}+T_{eq}}\right)\frac{r}{c} = p_{max} \tag{19}$$

which is the maximum level p_{max} of positive incentives that the defender would be willing to provide. In accordance with Proposition 5, p_{max} is inversely proportional to c. Again, inserting (7) and (8) into (19) gives

$$p_{max} = \begin{cases} (1 - \frac{Ar}{2aR} - \frac{Br}{2aR}(\frac{4bRa}{Bkr^2})^{\frac{k}{k-1}})\frac{r}{c} & \text{when } B > 0 \\ (1 - \frac{Ar}{2Ra})\frac{r}{c} & \text{when } B \leq 0 \end{cases} \tag{20}$$

Three outcomes are possible. First, if $p_{min} > p_{max}$, the terrorist would not be satisfied by the maximum level of positive incentives that the defender is willing to provide, and the equilibrium will be given by t_{eq}, n_{eq}, and T_{eq}. This means that the default game gets played, with no provision of positive incentives. Second, if $p_{min} \leq p_{max}$, the terrorist is satisfied by the level of positive incentives provided by the defender. The defender could in principle choose any value of p between p_{min} and p_{max}; however, we assume that the defender chooses the value p_{min} that makes the terrorist exactly indifferent between $T = 0$ and T_{eq}. This approach is common in the literature on principal-agent games (see for example Tirole 1989), where the principal makes the agent indifferent between participating or not. It is therefore called a participation constraint, and sometimes an individual-rationality constraint. The terrorist thus earns Cp_{min}. The defender, having assessed the situation and determined that positive incentives can be provided, eliminates the negative incentives by setting $n = 0$, chooses $t > 0$ arbitrarily small but positive to win the contest, and earns utility $u = r - cp_{min}$. (Although an arbitrarily small but positive $t > 0$ wins the contest mathematically in (1) when $T = 0$, in practice we can assume that both agents refrain from investing any effort; however, since both agents realize the defender's superiority, the defender wins the contest.) Third, if $p_{max} < 0$, the defender provides no positive incentives. Comparing (8), (9), (18), and (20), observe that T, U, p_{min}, and p_{max} equal zero for the same parameter values. This means that as p_{max} decreases toward zero, p_{min} decreases to zero too, and as T and U decrease to zero, positive incentives play no role.

Comparing (18) and (20) gives

$$p_{min} \leq p_{max} \quad \Leftrightarrow \quad \begin{cases} (1 - \frac{Ar}{2Ra} - \frac{Br}{2Ra}(\frac{4bRa}{Bkr^2})^{\frac{k}{k-1}})\frac{R}{C} \leq \frac{r}{c} & \text{when } B > 0 \\ (1 - \frac{Ar}{2Ra})\frac{R}{C} \leq \frac{r}{c} & \text{when } B \leq 0 \end{cases} \tag{21}$$

which gives a role for positive incentives when the inequalities are satisfied. Observe that we will have $p_{min} = p_{max} = 0$ in (18) and (20) when $T = U = 0$ in (8) and (9), since the expressions inside the brackets are equivalent. This means that as the terrorist becomes disadvantaged (e.g., by larger values of A or B), the terrorist's attack effort and utility in (8) and (9) will decrease toward zero, and the role for positive incentives will vanish (in the sense that p_{min} and p_{max} decrease toward zero).

Proposition 6 (1) *When* $p_{min} > p_{max}$, *the advantaged terrorist does not respond to affordable levels of positive incentives, so* $p = 0$. (2) *When* $p_{min} \le p_{max} \ge 0$, *the terrorist does respond to positive incentives, and* $p = p_{min}$. (3) *When* $p_{max} < 0$, *the defender provides no positive incentives to the disadvantaged terrorist,* $p = 0$.

Proof Follows from (8), (9), (18), (20), (21), (15) and (16). □

The first case in Proposition 6 expresses the idea that a sufficiently advantaged terrorist cannot be successfully deterred from attacking by positive incentives. This can arise when the terrorist is advantaged by $(1 - \frac{Ar}{2Ra} - \frac{Br}{2Ra}(\frac{4bRa}{Bkr^2})^{\frac{k}{k-1}})$ being large when $B > 0$, by $(1 - \frac{Ar}{2Ra})$ being large when $B \le 0$, and by R/C being large relative to r/c. For example, this can occur when the terrorist's unit costs A and B are low, or when the terrorist's asset valuation divided by its unit benefit of positive incentives is large relative to the corresponding ratio for the defender. The second case in Proposition 6 expresses the idea that a moderately advantaged terrorist can be successfully deterred from attacking by positive incentives. If the terrorist violates the trust associated with positive incentives, the defender sets $p = 0$, and chooses t and n in (7). The third case in Proposition 6 expresses the idea that a sufficiently disadvantaged terrorist, caused for example by high unit costs A or B, again provides no role for positive incentives.

Figures 3 and 4 show the five variables characterizing the equilibrium solution, and additionally p_{min}, p_{max}, and p, as functions of B for the same baseline parameter values as in the previous section ($a = b = A = B = 1$, $r = R = 4$, and $k = 0.5$); we also assume $c = 1$ and $C = 0.4$. Observe that $p_{min} \le p_{max}$ when $B > 0.63$, which provides a role for positive incentives. As B increases above $\sqrt{2}$, as in Fig. 2, the terrorist is deterred by the negative incentives, and positive incentives play no role.

The intermediate range $0.63 < B < \sqrt{2}$ in Figs. 3 and 4 (corresponding to the second case in Proposition 6) is extremely interesting. It means that positive incentives can play a role only when negative incentives have an intermediate impact on the terrorist. When negative incentives have too small of an impact ($B < 0.63$), the terrorist is too advantaged, and would require a great deal in the way of positive incentives to refrain from attacking, which would be overly costly for the defender. To see this, see the expression for B_D in (14);

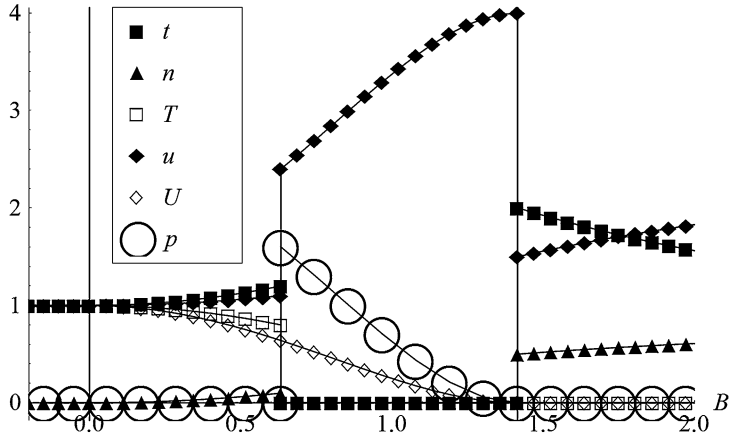

Fig. 3 Behavior of the equilibrium solution (including positive incentives) as a function of the incentive parameter B

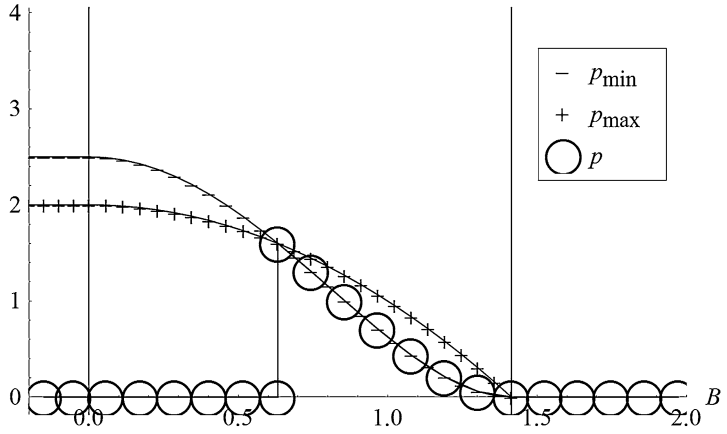

Fig. 4 p_{min}, p_{max} and p as functions of the incentive parameter B

essentially, the terrorist is advantaged because its asset valuation R is sufficiently large to make attacking highly attractive. Therefore, positive incentives of a magnitude acceptable to the defender will not be helpful in this case, and it is only moderately worthwhile for the defender to apply negative incentives. Conversely, when negative incentives have sufficiently high impact ($B > \sqrt{2}$), the terrorist is already disadvantaged and deterred by the negative incentives, and there is no need for positive incentives.

6 Conclusion

We develop a model where the defender controls an asset which is under attack by a terrorist. In addition to defending the asset, the defender can employ negative and positive incentives to impact the terrorist's attack. The model distinguishes between the effects of negative incentives ("sticks") and positive incentives ("carrots") for influencing the behavior of intelligent and adaptable adversaries. The defender chooses the levels of defensive effort and negative incentives in the first period of a two-period game. The terrorist then chooses the level of attack effort in the second period (with the terrorist's unit cost of attack being influenced by the negative incentives, if any). The defender also determines (by two conditionality tests) whether positive incentives can be applied. That is, the defender evaluates both games (one with the option of positive incentives, and one without), and plays whichever one is better (since the defender is the first mover). The parameters of both games are common knowledge. At equilibrium in the game with positive incentives, the defender offers the terrorist the smallest level of positive incentives in the first period required to prevent an attack in the second period. The game with positive incentives will be chosen over the game without positive incentives when the defender's unit cost of providing positive incentives is small compared with the terrorist's unit benefit of receiving positive incentives. This was the case, first, when PLO dismantled Black September, incurring modest costs while the earlier elite soldiers enjoyed positive incentives starting a more conventional life. Second, and similarly, the authorities in Northern Ireland incurred modest costs providing furloughs, while the imprisoned terrorists thereby enjoyed positive incentives through developing family ties. When the terrorist is extremely advantaged (i.e., with a low unit cost of

attack, taking into account any increases in unit cost due to negative incentives, or a high asset valuation), positive incentives cannot deter the terrorist from attacking, because the asset is too valuable to the terrorist relative to the cost of attacking. This was the case when the U.S. bombed Libya in 1986, forcing Libya to realize that the costs of continued terrorism are too high. However, a less advantaged terrorist can be successfully deterred from attacking by positive incentives, and this can also be worthwhile for the defender. (If the terrorist violates the trust associated with positive incentives, which may occur when compliance cannot be adequately monitored and enforced, the defender is assumed to eliminate the use of positive incentives.) Finally, when the terrorist is sufficiently disadvantaged, it is already deterred from attacking, and positive incentives play no role.

We demonstrate a complex relationship between negative and positive incentives, which we believe has not been shown previously. In particular, we find that positive incentives can play a role only when negative incentives have an intermediate impact on the terrorist. When negative incentives have too little impact, the terrorist is too advantaged for affordable levels of positive incentives to be effective at achieving deterrence. This accords with past observations by Pruitt (2006), suggesting that negotiations with terrorists are likely to be successful only when the goals of the terrorist group are relatively modest or pragmatic; by contrast, apocalyptic terrorists may be so "advantaged" by their fanatic devotion to their cause that the benefits of any positive incentives the defender might be willing to offer would pale by comparison. From this perspective, "appeasement" is not necessarily always undesirable, but will be so when the defender has underestimated the goals and/or devotion of the adversary. Conversely, when negative incentives have a sufficiently large impact, the terrorist is already disadvantaged and deterred by negative incentives alone, and there is no need for the defender to offer positive incentives.

Machiavelli (1492) recommended that the Prince should work to be both loved (positive incentives) and feared (negative incentives). He realized that the Prince could not always rely on being loved, and then should rely on fear. This paper proceeds beyond Machiavelli's qualitative suggestion. In particular, we recommend that the defender should indeed rely on negative incentives when they are likely to be sufficiently effective. (Of course, history has shown that negative incentives can sometimes be highly effective.) When the defender cannot expect negative incentives to be highly effective (e.g., due to possible perverse effects of counterterrorism), positive incentives can play a role. However, relying on positive rather than negative incentives to deter terrorists from attacking requires the development of trust on the part of the defender (based on the terrorist's recognition that the defender can monitor and enforce compliance). If the terrorist violates this trust, the defender would presumably refrain from offering positive incentives, and instead rely on negative incentives. Finally, negative incentives will in general be ineffective against a highly advantaged terrorist, but in that case, there is no role for positive incentives either, and the defender must rely exclusively on defending itself, rather than on deterring attacks.

In this work, the possibility of a perverse effect of counterterrorism (e.g., the emergence of hatred) among terrorists is modeled by a parameter that can decrease the terrorist's unit cost of attack when negative incentives are imposed, thereby decreasing their effectiveness. We show that the potential for perverse effects of counterterrorism can cause the defender to rely on positive incentives and decrease or eliminate its reliance on negative incentives at equilibrium.

A relevant policy issue is what the implications of our model are with regard to whether U.S. post-9/11 counterterrorism policy is on the right track. We feel that both economics as a discipline, and the U.S. as a society, are only in the early stages of being able to address this question. The intent of our analysis is not to answer the question of whether current policy

is justified, but to provide a framework within which the ongoing debate over the merits of that policy can be more productively continued. Right now, it often seems as if the debate over issues like whether the war in Iraq is making us safer from terrorism is being conducted on the basis of people's prior attitudes towards the use of violence before the war in Iraq had even started, with some people believing that the world is a nasty and brutish place (Hobbes 1651), so that violence is inevitable and necessary, and others believing that violence only begets more violence, and almost all problems have nonviolent solutions (Brock 1999).

In formulating this paper, we have found that we agree with Obama's (2009) Nobel Peace Prize Lecture: "...we will not eradicate violent conflict in our lifetimes. There will be times when nations—acting individually or in concert—will find the use of force not only necessary but morally justified." Thus, the intent of this paper is to move the discussion in the direction of more rigorous analysis, by indicating what types of questions would need to be answered in order to reach a consensus on the circumstances under which use of force will be beneficial, rather than the current political impasse brought about by conflicting prior opinions regarding the use of violence in general. In this respect, we view the current status of the "global war on terrorism" as analogous to the early days of the Cold War. Just as policymakers in the U.S. were originally unsure whether threatening to retaliate against a nuclear strike would make us safer or less safe from nuclear war, before work such as by Schelling (1960, 1978) and Aumann and Maschler (1995) forged a consensus that mutually assured destruction was in fact a stable equilibrium strategy, so too policymakers and academics today are in need of a rigorous way to answer questions such as whether the war in Iraq is making the U.S. safer or less safe from terrorism.

Before such questions can be answered, however, further research is required. Achieving a more complete and thorough understanding of possible perverse effects of counterterrorism will require not only empirical work on the nature of particular terrorist groups, but also further theoretical development. In particular, we believe that an important next step is to understand the conditions under which a defender's threat of retaliation or other enforcement mechanism is credible, if an attacker attempts to launch an attack after accepting positive incentives as a reward for not attacking. In addition, it will obviously be critical to determine how the desirability of using negative incentives depends on the extent of uncertainty about key parameters of our model, since it will typically not be known for sure in advance whether negative incentives will lead to perverse effects.

Acknowledgements We thank two anonymous reviewers of this journal for useful comments.

Appendix: Dependence of t, n, T, u, and U on k

To illustrate Proposition 2, consider how the interior solution in (7)–(9) change with k when $B > 0$. Differentiation gives

$$\frac{\partial t}{\partial k} = -\frac{Br^2\left(\frac{4abR}{Bkr^2}\right)^{\frac{k}{k-1}}\left(k + \ln\left(\frac{4abR}{Bkr^2}\right) - 1\right)}{4a^2(k-1)^2R} \tag{A.1}$$

$$\frac{\partial n}{\partial k} = -\frac{\left(\frac{4abR}{Bkr^2}\right)^{\frac{1}{k-1}}\left(k + k\ln\left(\frac{4abR}{Bkr^2}\right) - 1\right)}{(k-1)^2k} \tag{A.2}$$

$$\frac{\partial T}{\partial k} = \frac{Br^2\left(\frac{4abR}{Bkr^2}\right)^{\frac{k}{k-1}}\left(k + \ln\left(\frac{4abR}{Bkr^2}\right) - 1\right)}{4a^2(k-1)^2R} \tag{A.3}$$

$$\frac{\partial u}{\partial k} = \frac{b(\frac{4abR}{Bkr^2})^{\frac{1}{k-1}} \ln(\frac{4abR}{Bkr^2})}{(k-1)k} \tag{A.4}$$

$$\frac{\partial U}{\partial k} = -\frac{2^{\frac{k+1}{k-1}} Br(\frac{abR}{Bkr^2})^{\frac{k}{k-1}} (Br(\frac{4abR}{Bkr^2})^{\frac{k}{k-1}} + Ar - 2aR)(k + \ln(\frac{4abR}{Bkr^2}) - 1)}{a^2(k-1)^2 R} \tag{A.5}$$

References

Aumann, R., & Maschler, M. (1995). *Repeated games with incomplete information*. Cambridge: MIT Press.

Bandyopadhyay, S., & Sandler, T. (2011). The interplay between preemptive and defensive counterterrorism measures: A two-stage game. *Economica*. doi:10.1111/j.1468-0335.2009.00823.x.

Bier, V. M., Oliveros, S., & Samuelson, L. (2007). Choosing what to protect: strategic defensive allocation against an unknown attacker. *Journal of Public Economic Theory*, *9*, 563–587.

Brock, P. (1999). *Varieties of pacifism: A survey from antiquity to the outset of the twentieth century*. New York: Syracuse University Press.

Burton, F. (2008). *Ghost: confessions of a counterterrorism agent*. New York: Random House.

Dalvi, N., Domingos, P., Mausam, M., Sanghai, S., & Verma, D. (2004). Adversarial classification. In *Proceedings of the 2004 ACM SIGKDD international conference on knowledge discovery and data mining*, Seattle, WA (pp. 99–108).

Dirks, T. (1969), Review of "Butch Cassidy and the Sundance Kid (1969)", http://www.filmsite.org/butc3.html.

Enders, W., & Sandler, T. (2004). What do we know about the substitution effect in transnational terrorism. In A. Silke & G. Ilardi (Eds.), *Research on terrorism: trends, achievements and failures*. London: Frank Cass.

Frey, B. S. (2004). *Dealing with terrorism: stick or carrot*. Northampton: Edward Elgar.

Frey, B. S., & Luechinger, S. (2003). How to fight terrorism: alternatives to deterrence. *Defense and Peace Economics*, *14*, 237–249.

Ganor, B. (2005). *The counter-terrorism puzzle: a guide for decision makers*. New Brunswick: Transaction Publishers.

Glaeser, E. L. (2005). The political economy of hatred. *Quarterly Journal of Economics*, *120*, 45–86.

Hausken, K. (2006). Income, interdependence, and substitution effects affecting incentives for security investment. *Journal of Accounting and Public Policy*, *25*, 629–665.

Hobbes, T. (1651). *Leviathan*. London: Dent.

Hoffman, B. (1998). *Inside terrorism*. New York: Columbia University Press.

Hoffman, B. (2001), All you need is love, The Atlantic, http://www.theatlantic.com/doc/200112/hoffman.

Hirshleifer, J. (1995). Anarchy and its breakdown. *Journal of Political Economy*, *103*, 26–52.

Keohane, N., & Zeckhauser, R. J. (2003). The ecology of terror defense. *Journal of Risk and Uncertainty*, *26*, 201–229.

Lakdawalla, D., & Zanjani, G. (2005). Insurance, self-protection, and the economics of terrorism. *Journal of Public Economics*, *89*, 1891–1905.

Lutfi, A. (2001). Blowback: China and the Afghan Arabs. *Issues and Studies*, *37*(1), 160–214.

Obama, B. H. 2009. The Nobel peace prize lecture, Oslo, December 10, 2009, http://nobelprize.org/nobel_prizes/peace/laureates/2009/obama-lecture_en.html.

Machiavelli, N. (1492). *The prince*. Cambridge: Cambridge University Press.

Pruitt, D. G. (2006). Negotiation with terrorists. *International Negotiation*, *11*(2), 371–394.

Salop, S. C., & Scheffman, D. T. (1983). Raising rivals' costs. *American Economic Review*, *73*, 267–271.

Sandler, T., Arce, D. G., & Enders, W. (2008). Copenhagen consensus 2008 challenge paper: terrorism.

Schelling, T. (1960). *The strategy of conflict*. Cambridge: Harvard University Press.

Schelling, T. (1978). *Micromotives and macrobehavior*. New York: Norton.

Skaperdas, S. (1996). Contest success functions. *Economic Theory*, *7*, 283–290.

Spector, B. I. (1998). Deciding to negotiate with villains. *Negotiation Journal*, *14*(1), 43–59.

Spector, B. I. (2003). Negotiating with villains revisited: research note. *International Negotiation*, *8*(3), 616–626.

Tzu, S. (320 Before Christ). *The art of war*. Translated by Griffith, S.B. (1963). London: Oxford University Press.

Tirole, J. (1989). *The theory of industrial organization*. New York: MIT Press.

Tullock, G. (1967). The welfare costs of tariffs, monopolies, and theft. *Western Economic Journal*, *5*, 224–232.