Defensive weapons and star wars: a supergame with optimal punishments

Defensive Weapons and Star Wars: A Supergame with Optimal Punishments[1]

Giampiero Giacomello[§] and Luca Lambertini[#]

§ Department of Political and Social Studies, University of Bologna

Strada Maggiore 45, 40125 Bologna, Italy

giampiero.giacomello@unibo.it

Corresponding author

Department of Economics, University of Bologna

Strada Maggiore 45, 40125 Bologna, Italy

luca.lambertini@unibo.it

March 2, 2015

[1]We would like to thank the Editor, Christos Kollias, an anonymous referee, Erik Jones, Arsen Palestini, Alessandro Pascolini, Mike Plummer and Alessandro Tampieri for useful comments and suggestions. The usual disclaimer applies.

Abstract

We model the perspective faced by nuclear powers involved in a supergame where nuclear deterrence is used to stabilise peace. This setting allows us to investigate the bearings of defensive weapons on the effectiveness of deterrence and peace stability, relying on one-shot optimal punishments. We find that the sustainability of peace is unaffected by defensive shields if the latter are symmetric across countries, while asymmetric endowments of such weapons have clearcut destabilising consequences.

Keywords: defensive shields, deterrence, wargames, balance of power

JEL codes: C73, F51

1 Introduction

The long-standing debate on the relationship between the distribution of power and stability (or the outbreak of war) can be summarised in two polar positions according to which stability is the outcome of either (i) the balance of power or (ii) the preponderance of power (for an account, see Powell, 1996, 1999, among many others). A relevant aspect of this issue is that of the distribution of defensive weapons across countries, as an uneven distribution of such endowments might trigger preemptive unilateral strikes.[1] Here, we intend to nest ourselves in the motherlode of this discussion, by assessing the bearings of the worldwide distribution of defensive weapons on stability, as a specific component of the *balance vs unbalance of power* debate.

In two recent papers, Chassang and Padró I Miquel (2009, 2010) discuss the role of defensive shields in nuclear conflicts, an issue that has remained a hot one throughout the Cold War and has become even more so in coincidence with the *Strategic Defense Initiative* (commonly known as 'Star Wars') during the Reagan administration and afterwards. This particular topic nests into a more general debate concerning escalation and deterrence, precisely because defensive shields may ultimately jeopardise the deterrence capability of nuclear arsenals.[2] Consequently, a specific discussion has been carried out on this aspect, with a striking lack of consensus. Some, like Wilkening (2000) and Lebovic (2002) have focussed on the efficiency of ballistic missile defence, while others have reached opposite conclusions as to the interplay between defensive shields and deterrence: on one side, there are scholars convinced of the destabilising effects of shields (e.g., Brams and

[1] On preemptive attacks generated by commitment mechanisms, see Powell (2006).

[2] The number of contributions being published in recent years demonstrate that these aspects remain at the core of international politics even after the end of the Cold War (see Powell, 1990; Zagare and Kilgour, 2000; Danilovic, 2002; Freedman, 2004; Zagare, 2004, to mention only a few).

Kilgour, 1988; and Miller, 2001); on the other, there are those sustaining the opposite view (e.g., Powell, 1990, 2003; and Lindsay and O'Hanlon, 2001). In summary, while classical deterrence theory usually relies on a chicken game structure in which backing down is preferred to the all-out conflict, and is affected by logical inconsistencies (generating equilibria which sometimes are not subgame perfect, as in Brams and Kilgour, 1988), perfect deterrence theory (see Zagare and Kilgour, 2000) holds that countries my prefer symmetric warfare to backing down unilaterally, so that the underlying game structure is not a chicken but rather looks like a prisoners' dilemma. In general, the literature generated by perfect deterrence theory uses a multistage extensive form game where asymmetric escalation is investigated, and there emerges that the unilateral adoption of a defensive shield enhances deterrence but may leave challegers dissatisfied with the *status quo* (Zagare and Kilgour, 1993, 2000; Quackenbush, 2006; Quackenbush and Drury, 2011).

Our standpoint is that additional insights on the pros and cons of defensive shields could be grasped by embedding the constituent game into a long run perspective using the toolkit of repeated game theory. Chassang and Padró I Miquel (2009, 2010) use a repeated game framework based on a prisoner's dilemma to show that "the unilateral acquisition of defensive weapons raises the stronger country's predatory incentives and reduces the sustainability of peace" (Chassang and Padró I Miquel, 2009, p. 282). To do so, they model the supergame following the route traced by Friedman (1971), whereby trigger strategies imply that a unilateral strike is followed by an everlasting Nash punishment consisting in permanent conflict.[3] This particular component of the supergame - the design of the punishment phase

[3]Using the supergame as a benchmark, Chassang and Padrò I Miquel (2009) investigate the nature of the equilibium using the risk dominance refinement (Harsanyi and Selten, 1988), while in the subsequent paper they build up a global game (Carlsson and van Damme, 1993).

- is indeed critical, as perpetual conflict may not be sustainable if countries have conspicuous endowments of nuclear weapons, whose massive use in the very first phase of the confrontation would very rapidly annihilate contenders altogether, leaving possibly open the way for a retaliation strike but almost surely not for an infinite repetition of the constituent game.[4] Therefore, we propose here an alternative approach to the construction of the supergame, based on one-shot optimal punishment (Abreu, 1986). This type of punishment, besides being the most efficient one, has the additional advantage - which is particularly convenient when it comes to assessing the effectiveness of deterrence in preventing the emergence of nuclear conflicts - of allowing us to do altogether away with the Nash equilibrium of the constituent game in building up the equilibrium conditions based on the players' intertemporal payoffs. Additionally, our setup is modelled in such a way to allow for the fully analytical characterisation of optimal one-shot punishment in the asymmetric case where only one country is endowed with defensive shields, and this - to the best of our knowledge - is a rare property in the extant applications of optimal penal codes, where any asymmetry across players is known to usually prevent the attainment of fully analytical solutions.

As in Chassang and Padró I Miquel (2009, 2010), we examine a two-country game, whose constituent stage is a prisoners' dilemma. The main results yielded by our setup can be summarised as follows. In the symmetric case where both countries have defensive shields, (i) the stability condition for the sustainability of perpetual peace is completely independent of the shields,

[4]In this respect, it is worth recalling that, during the Cuba crisis, President Kennedy ordered (among other things) to keep aloft the Strategic Air Command bombers armed with nuclear weapons for a gross total of 7000 megatons around the clock. It is hardly credible that an infinitely repeated game could follow the very first stage. Also note that, for several years, General Curtis LeMay was convinced that his strategic bombers could 'bring the enemy back to the stone age in the first six hours', this all-out attack being known as the 'Sunday punch' (for more details on this, see Rhodes, 1995).

while (ii) the one-shot optimal punishment is inversely related to the shield's effectiveness. If instead countries are endowed with asymmetric amounts of defensive weapons, or their effectiveness is asymmetric, then the resulting stability conditions are asymmetric and, more importantly, conducive to less stable outcomes as compared to the fully symmetric setting. This is due to the fact that a country's critical discount factor above which perpetual peace is preferable to war is increasing in the effectiveness of the shield held by the country itself, for which the temptation to strike first becomes therefore harder to resist.

The main conclusion reached by our analysis - namely, that symmetry in defensive shields favours stability, and conversely asymmetry is destabilising - may seem at odd not only with a large part of the discussion on the matter which has taken place in recent literature but also with major historical episodes in which large conflicts have taken place under symmetric conditions, up to World War II. However, all of these conflicts (i) broke out before the nuclear age, and (ii) were fought with weapons which could not achieve the mass destruction nuclear weapons are capable of (with the notable exception of the conclusion of WWI, of course). During the early phase of the Cold War, well before Reagan's Star Wars project started, the slightest possibility that a single Soviet nuclear weapon could hit the US metropolitan territory was sufficient for the White House to put aside any temptation to strike first, the bold attitude of Strategic Air Command notwithstanding (cf. Rhodes, 1995, among others). The unilateral creation of defensive shields may legitimately appear as conducive to war, as the adopter of such a shield could be tempted to rely on it and strike first. That is, asymmetry in offensive weapons intrinsically differs from asymmetry in defensive instruments, as if the latter is sufficiently pronouced, a country ahead of others along this dimension may in fact use offensive weapons it wouldn't use otherwise, irrespective of the relative endowments of the offensive weapons themselves.

This must be considered in combination with the long run perspective of survival vs annihilation, which was not associated with the pre-nuclear age, and makes the interplay between intertemporal prefences and defensive shields relevant, for any positive endowments of offensive weapons available to the countries involved.

The remainder of the paper is organised as follows. The basic model is illustrated in section 2. Section 3 investigates the symmetric supergame. The consequences of asymmetry are outlined in section 4. Concluding remarks are in section 5.

2 The setup

The model can be quickly laid out as follows. Players are two countries, labelled as 1 and 2, interacting over time $t = 0, 1, 2, 3...\infty$. Country i's per-period payoff is

$$u_i = \pi \left(1 - a_j \left(1 - \sigma\right) + z a_i \left(1 - \sigma\right)\right) - b a_i^2 - C \left(\sigma\right), \tag{1}$$

$i, j = 1, 2$, $j \neq i$. In (1),

- $\pi > 0$ is the instantaneous welfare associated with perpetual peace, gross of any costs associated with building up offensive weapons a_i and the defensive shield $\sigma \in [0, 1]$. The function $C\left(\sigma\right)$ measures the cost of building up the shield. Given that we are dealing with the analysis of a possible nuclear confrontation, net payoffs are not necessarily positive (indeed they may become strongly negative). Even in a peaceful situation, the investment in a defensive shield may be so costly to cause $\pi < C\left(\sigma\right)$.

- The term $1 - a_j \left(1 - \sigma\right) + z a_i \left(1 - \sigma\right)$ multiplying π determines how much the aforementioned utility is modified by a war where the intensity of

attacks are a_i and a_j, respectively, and both countries are endowed with a shield σ. Note that to be fully effective the shield should have size one. Accordingly, to capture the idea that, reasonably, building up the 'perfect shield' is indeed an unrealistic objective, we may assume the cost function $C(\sigma)$ associated with the R&D activities and the implementation of the defensive shield obeys the following:[5]

$$C(\sigma) = 0;\ \frac{\partial C(\sigma)}{\partial \sigma} > 0;\ \frac{\partial^2 C(\sigma)}{\partial \sigma^2} > 0;\ \lim_{\sigma \to 1} C(\sigma) = \infty. \tag{2}$$

Parameter $z > 0$ measures how sensitive country i is to a strike of intensity a_i brought to bear against country j, as opposed to the damages inflicted by the opponent through an attack of size a_j. For any $z \in (0,1)$, country i exhibits what we may define as a net damage aversion, in that any advantage generated by its own strike against j is offset, all else equal, by the damage caused by a comparable amount of, say, nuclear bombing borne by i itself. If instead $z > 1$, country i is keen on attacking the adversary, no matter what (this may reflect the fact that hawks are in control of i's actions).[6]

- ba_i^2 is the production cost associated with building up the amount of

[5] For instance, a cost function with these properties is

$$C(\sigma) = \frac{1}{1-\sigma} - 1.$$

This would indeed imply $\pi - C(\sigma) < 0$ for values of σ sufficiently close to one.

[6] Posing $z > 1$ one could interpret this parameter as the expected value of a multiplicative shock. More generally, one could set

$$u_i = \pi\left(1 - \frac{a_j}{\theta_j}(1-\sigma) + \frac{a_i}{\theta_i}(1-\sigma)\right) - ba_i^2 - C(\sigma)$$

with θ_i and θ_j being stochastic variables with expected values $E(1/\theta_i) = z_i > 1$ by Jensen's inequality. This would affect the levels of critical discount factors in the repeated games examined below without, however, interacting significantly with the efficiency of the shields. For more on this type of uncertainty, see Klemperer and Meyer (1986).

nuclear weapons a_i, at decreasing returns to scale.

It is also worth noting that (1) is defined in such a way that each country values the effect of both shields' effectiveness in intercepting the enemy weapons. Alternatively, one could assume that only one's own shield's effectiveness does matter, which might improve the effectiveness of deterrence (see Quackenbush, 2006, p. 537).

Before delving into the structure and properties of the repeated game, we briefly illustrate the basic features of the constituent one-shot game.

2.1 The one-shot game

The equilibrium of the one-shot constituent game can be quickly characterised. The first order condition of country i is:

$$\frac{\partial u_i}{\partial a_i} = z\left(1-\sigma\right)\pi - 2ba_i = 0 \tag{3}$$

yielding $a_i^* = z\left(1-\sigma\right)\pi/\left(2b\right)$. For future reference, it is worth noting that (3) implies that the two countries' best reply functions are orthogonal to each other, so that the resulting equilibrium is the intersection of dominant strategies. The associated utility is

$$u^* = \frac{4b\left[\pi - C\left(\sigma\right)\right] - z\left(2-z\right)\left(1-\sigma\right)^2\pi^2}{4b} > 0 \tag{4}$$

for all

$$b > \max\left\{0, \frac{z\left(2-z\right)\left(1-\sigma\right)^2\pi^2}{4\left[\pi - C\left(\sigma\right)\right]}\right\}. \tag{5}$$

To this regard, note that, for all $z \in (0,2)$ any $b > 0$ suffices to ensure that $u_i^* > 0$.

There remains to verify that $z \in (0,2)$ suffices to imply that the underlying structure corresponds to a prisoners' dilemma. To see this, observe that

the decision whether to attack or not can be evaluated by looking at Matrix 1. Pure strategies P and A stand for *remaining peaceful* and *attacking*, respectively. The outcome (A, A) portrays the equilibrium of the game in continuous strategies described above.

		2	
		P	A
1	P	u^{PP} ; u^{PP}	u^{PA} ; u^{AP}
	A	u^{AP} ; u^{PA}	u^{*} ; u^{*}

Matrix 1: The one-shot game in reduced form

The payoff accruing to each player when both choose P is generated by $a_i = a_j = 0$, so that it corresponds to

$$u^{PP} = \pi - C(\sigma) \tag{6}$$

with

$$u^{PP} - u^{*} = \frac{z(2-z)(1-\sigma)^2 \pi^2}{4b} > 0 \tag{7}$$

for all $z \in (0, 2)$.

Then, consider the unilateral deviation from the peaceful outcome via the dominant strategy $a_i^{*} = z(1-\sigma)\pi/(2b)$, yielding to the deviator a payoff equal to

$$u^{AP} = \pi - C(\sigma) + \frac{z^2(1-\sigma)^2 \pi^2}{4b} > 0 \tag{8}$$

while the country enduring the strike gets

$$u^{PA} = \pi - C(\sigma) - \frac{z(1-\sigma)^2 \pi^2}{2b} \tag{9}$$

with

$$u^{AP} - u^{PP} = u^* - u^{PA} = \frac{z^2 (1-\sigma)^2 \pi^2}{4b} > 0 \text{ always.} \tag{10}$$

Hence,

Lemma 1 *The one-shot game reflects the structure of a prisoners' dilemma for all* $z \in (0, 2)$.

Modelling the constituent game as a prisoners' dilemma is in line with Chassang and Padró I Miquel (2009, p. 283; and 2010, p. 1826), although not with a large literature on deterrence that has been flourishing after Schelling (1960), where the chicken game is taken as the reference structure to describe strategic interaction (see, among many others, Brams and Kilgour, 1988; Powell, 1990; Zagare and Kilgour, 2000). It is worth observing, however, that the parametric condition highlighted in the above Lemma is needed to establish that peace Pareto-dominates conflict - which must be the case in order for countries' preference structure to be sensible - and has nothing to do with the comparison between the payoffs associated with the asymmetric outcomes and those generated by war. This amounts to saying that the reduced-form game in Matrix 1 reflects a prisoners' dilemma as an endogenous results of objective functions (1), combined with a reasonable assumption concerning the relative size of u^{PP} and u^*. In other words, preferences (1) generate e reduced-form game where states prefer an all-out war to backing down,[7] which is itself is a stance in line with perfect deterrence theory, as stressed by Zagare and Kilgour (2000). As a consequence, the emergence of asymmetric equilibria along the secondary diagonal and the associated

[7] This is particularly relevant if the unilateral adoption of strategy A implies, as in the game under consideration, an actual strike and not simply the adoption of an aggressive stance. The overall (political, economic, environmental and demographic) consequences of withstanding a nuclear attack without responding might indeed be considered as worse than those generated by symmetric warfare.

strategic uncertainty discussed in Zagare and Kilgour (2000, p. 20) are ruled out, and we are left with a unique equilibrium of the one-shot game, where each country chooses to attack the rival. Hence, deterrence comes into the picture via the repetition of the constituent game, rather than in a multistage game like those discussed in Brams and Kilgour (1988), Zagare and Kilgour (1993, 2000), Powell (2003) as well as Quackenbush (2006), where both a response-in-kind and the second-strike capability play a role. The prisoners' dilemma structure and the full symmetry of the constituent game are both aligned with the view, adopted by Zagare and Kilgour (2000) in structuring their perfect deterrence theory, that the presence of a unique equilibrium at the intersection of dominant strategies can be most conducive to deterrence stability.

As is well known, repeating the prisoners' dilemma over an infinite horizon may allow players to attain Pareto-superior outcomes (as compared to the Nash equilibrium of the one-shot constituent game), provided they are enough forward looking, i.e., their intertemporal preferences must satisfy a condition whose specific nature will depend upon the rules of the supergame itself (in particular, the duration and intensity of the punishment phase). These aspects are investigated in the next section.

3 The supergame

The backbone of the literature on supergames is the so called *Folk Theorem*, that has evolved over the decades taking several alternative forms. From a qualitative point of view, the different formulations of this theorem yield the invariant message that infinite repetition allows players to reach any equilibrium outcomes Pareto-dominating that of the one-shot game, provided the players themselves be patient enough. The research efforts carried out over the last four decades have pointed at refining the structure of the punishment

phase in order to yield the most efficient outcome, all else equal. For this reason, we will focus here on a supergame based upon Abreu's one-shot optimal punishments (Abreu, 1986). Then, the performance of optimal punishments will also be contrasted with that of the version of this supergame based on an infinite Nash reversion, as in Friedman (1971), although the latter has to be taken with a pinch of salt for the aforementioned reasons.

Countries, being symmetric in all respects, are assumed to have the same time preferences, represented by the time-invariant discount factor $\delta \in [0,1]$. The rules of the supergame unravelling under the deterrence exerted by one-shot optimal punishments can be spelled out as follows.

- At $t = 0$, both countries play strategy P (i.e., $a_i = 0$).
- At any $t \geq 1$, both countries keep playing P provided both have played P at $t-1$; otherwise, if any deviation from $(0,0)$ has been observed at $t-1$, then at t countries are to implement the symmetric optimal punishment a^{op}.
- Subsequently, at $t+1$ players return to strategy P provided that both of them have implemented the optimal punishment at t. If not (i.e., if at least one of them has deviated from the punishment), they are again required to play a^{op}.

This last rules entails that abiding by the symmetric optimal punishment simultaneously confines the punishment phase to a single period, while any deviation from it extends the punishment phase itself.

The individual payoff generated by the adoption of symmetric optimal punishments is:

$$u^{op} = \pi - C(\sigma) - a^{op}\left[ba^{op} + (1-z)(1-\sigma)\pi\right] \tag{11}$$

while the optimal unilateral deviation from the optimal penal code is given by the dominant strategy $a_i^* = z(1-\sigma)\pi/(2b)$, yielding the payoff

$$u^{Dop} = \pi - C(\sigma) - a^{op}(1-\sigma)\pi + \frac{z^2(1-\sigma)^2\pi^2}{4b} > u^{op} \text{ always.} \tag{12}$$

Perpetual peace is sustainable forever on the basis of the deterrence exerted by optimal punishments if and only if the following conditions are met:

$$u^{AP} - u^{PP} \leq \delta\left(u^{PP} - u^{op}\right) \tag{13}$$

$$u^{Dop} - u^{op} \leq \delta\left(u^{PP} - u^{op}\right) \tag{14}$$

Inequality (13) has to be met in order for unilateral deviations from the perpetual peace path not to be advantageous, while (14) establishes that deviations from the symmetric optimal punishment are not attractive.

This yields a system of two inequalities in two unknowns, the discount factor δ and the optimal punishment a^{op}. Solving, we obtain that (13-14) are simultaneously satisfied by any pair

$$a^{op} \geq \frac{z(1-\sigma)\pi}{b}\,;\ \delta \geq \frac{z}{4} \tag{15}$$

This proves the following result:

Lemma 2 *Under the threat of optimal punishments $a^{op} \geq z(1-\sigma)\pi/b$, the two countries can sustain perpetual peace provided their common time preferences satisfy $\delta \geq z/4$.*

A direct consequence of the above lemma is

Proposition 3 *The intensity of the optimal punishment is monotonically decreasing in the effectiveness of the shield σ. Instead, the stability of perpetual peace in altogether independent of σ.*

As an ancillary observation, it is worth noting that

$$\frac{z}{4} \in \left[0, \frac{1}{2}\right] \forall z \in [0, 2]\,. \tag{16}$$

More interesting is the fact that, the critical level of the discount factor being independent of the presence and size of shields that in turn affects the harshness of the punishment, the foregoing analysis also entails:

Corollary 4 *In the absence of shields (i.e., if $\sigma = 0$), perpetual peace is again sustainable for all $\delta \geq z/4$. However, the punishment has to reach its peak at $a^{op} \geq z\pi/b$.*

That is, the stability condition is exactly the same but it needs to be accompanied by a threat whose intensity is magnified as much as possible by the absence of any defensive screens whatsoever.

To appreciate the role of optimal punishments in determining the condition for peace stability, we may briefly dwell upon the stability condition generated by the alternative punishment consisting in an infinite reversion to the Nash equilibrium of the constituent game - as it has been used in the supergame analysed by Chassang and Padró I Miquel (2009). Like them, we will look at the admittedly quite unrealistic perspective in which perpetual war after a unilateral first strike is admissible. Following Friedman (1971), the perpetual stability of peace requires

$$\frac{u^{PP}}{1-\delta} \geq u^{AP} + \frac{\delta u^*}{1-\delta} \tag{17}$$

$$\Leftrightarrow \frac{z\left(1-\sigma\right)^2 \pi^2 \left(2\delta - z\right)}{4\left(1-\delta\right) b} \geq 0$$

which is met by all $\delta \geq z/2$. As expected, this of course is higher than (indeed, twice as high as) the critical discount factor generated by optimal

punishments.[8] It is worth stressing that, as is the case under optimal punishments, also the infinite Nash reversion yields a stability condition that is completely independent of defensive paraphernalia.

To this regard, it can be shown that completely different considerations hold true in the more realistic case in which war cannot continue *ad infinitum*. To see this, assume first that one period of symmetric conflict suffices to erase the two countries.[9] Peace stability requires:

$$\frac{u^{PP}}{1-\delta} \geq u^{AP} + \delta u^{*}. \tag{18}$$

In general, if war may be sustained for, say, $T-1$ periods after the first strike, the above inequality rewrites as

$$\frac{u^{PP}}{1-\delta} \geq u^{AP} + u^{*} \sum_{t=1}^{T} \delta^{t} \tag{19}$$

with

$$\sum_{t=1}^{T} \delta^{t} = \frac{1-\delta^{T+1}}{1-\delta} - 1 = \frac{\delta - \delta^{T+1}}{1-\delta}. \tag{20}$$

The presence of shields may extend the terminal date T ensuring the survival of both countries over a longer horizon, whereby we may take T as a function of shields σ and suppose that $\partial T(\sigma,\sigma)/\partial\sigma > 0$. If so, then

$$\frac{\partial\left(\dfrac{\delta - \delta^{T(\sigma,\sigma)+1}}{1-\delta}\right)}{\partial\sigma} = -\frac{\delta^{T(\sigma,\sigma)+1}\ln(\delta)\cdot\partial T(\sigma,\sigma)/\partial\sigma}{1-\delta} > 0 \tag{21}$$

because $\ln(\delta) < 0$ as δ, in general, will be lower than one. Hence, we may expect shields to exert a destabilising effect precisely because they are likely

[8] One could perform the same exercise using Axelrod's (1981) *tit-for-tat* strategies, with qualitatively analogous results.

[9] This is a realistic scenario, if one duly takes into account the global consequences of even a limited use of nuclear weapons. See, e.g., Mills *et al.* (2008).

to keep the supergame going after a first strike. This entails that assessing peace sustainability in a supergame relying on Nash punishments is a tricky task, as the answer ultimately depends on whether or not the punishment phase is indeed infinitely long.

The most relevant question arising from the foregoing analysis is whether the main conclusion, namely, that the critical threshold of time preferences is independent of the level of defensive technology, is robust to a relaxation of the assumption of full symmetry between countries. This issue is dealt with in the following section, where the model is extended to account for the more realistic possibility that defensive weapon endowments be asymmetric.

4 The asymmetric supergame

What if the countries' endowments of defensive weapons are asymmetric? Obviously this scenario includes, as an extreme case, the possibility that only one country has a defensive shield. In this regard, some of the extant literature argues that partial or extreme asymmetry is destabilising (Bundy et al, 1984/1985; Miller, 2001; Chassang and Padró I Miquel, 2009, 2010; Quackenbush and Drury, 2011, *inter alia*).

To investigate this issue in our repeated game setup, we proceed as follows. Country i's utility function is defined as follows:

$$u_i = \pi \left(1 - a_j \left(1 - \sigma_i\right) + z a_i \left(1 - \sigma_j\right)\right) - b a_i^2 - C\left(\sigma_i\right) \tag{22}$$

with $\sigma_i \geq$ and $\sigma_j \geq 0$ but, in general, $\sigma_i \neq \sigma_j$. The only element that remains symmetric is the value attached to peace, π. On the basis of (22), we can identify the relevant payoffs for the construction of the incentive compatibility constraints that have to be satisfied here in order for perpetual peace to be stable all over the supergame. Let's investigate optimal punishments first.

If $a_i = a_j = 0$, we have $u_i^{PP} = \pi - C(\sigma_i)$, while unilateral deviations from the peaceful path yield, respectively:

$$u_i^{AP} = \pi + \frac{(1-\sigma_j)^2 z^2 \pi^2}{4b} - C(\sigma_i) \ . \tag{23}$$

The punishment profits are

$$u_i^{op} = \pi\left(1 - a_j^{op}(1-\sigma_i) + z a_i^{op}(1-\sigma_j)\right) - b\left(a_i^{op}\right)^2 - C(\sigma_i) \tag{24}$$

where we may expect to observe $a_i^{op} \neq a_j^{op}$ given the asymmetry of the present setup. Likewise, deviations from the optimal penal code yield asymmetric payoffs:

$$u_i^{Dop} = \pi\left(1 + \frac{(1-\sigma_j) z^2 \pi}{4b} - a_j^{op}(1-\sigma_i)\right) - C(\sigma_i). \tag{25}$$

The set of constraints to be satisfied for the sustainability of perpetual peace is

$$\begin{array}{l} u_i^{AP} - u_i^{PP} \leq \delta_i\left(u_i^{PP} - u_i^{op}\right) \\ u_i^{Dop} - u_i^{op} \leq \delta_i\left(u_i^{PP} - u_i^{op}\right) \\ u_j^{AP} - u_j^{PP} \leq \delta_j\left(u_j^{PP} - u_j^{op}\right) \\ u_j^{Dop} - u_j^{op} \leq \delta_j\left(u_j^{PP} - u_j^{op}\right) \end{array} \tag{26}$$

with $i, j = 1, 2$, $i \neq j$. Note that the discount factor is indexed as well since under asymmetric conditions the relative stability thresholds are necessarily asymmetric. Solving these four inequalities, we obtain[10]

$$\delta_i \geq \delta_i^{op}(\sigma_i, \sigma_j) \equiv \frac{z(1-\sigma_j)^2}{4(1-\sigma_i)^2} \, ; \, a_i^{op} \geq \frac{z(1-\sigma_j)\pi}{b}. \tag{27}$$

Now note that

$$\frac{\partial \delta_i^{op}(\sigma_i, \sigma_j)}{\partial \sigma_i} = \frac{z(1-\sigma_j)^2}{2(1-\sigma_i)^3} > 0 \, \forall \, \sigma_i, \sigma_j \in [0, 1), \tag{28}$$

[10] The arguments in $\delta(\sigma, 0)$ indicate that i has a shield, while j hasn't. The opposite applies to $\delta(0, \sigma)$.

which of course entails that, for any given σ_j, an increase in the efficiency of country i's shield has a clearcut destabilising effect, this being fully in line with the analysis carries out on the basis of the symmetric model used in the previous section. However, the cross effect illustrated by the following partial derivative significantly modifies our appraisal of the problem at hand:

$$\frac{\partial \delta_i^{op}(\sigma_i, \sigma_j)}{\partial \sigma_j} = -\frac{z(1-\sigma_j)}{2(1-\sigma_i)^2} < 0 \forall \sigma_i, \sigma_j \in [0,1). \tag{29}$$

Intuitively, this shows that the deterrence exerted by any increase in country j's shield lowers the critical discount factor above which country i's intertemporal incentives are indeed peaceful ones. Using (28-29), we can build up the total differential

$$d\delta_i^{op}(\sigma_i, \sigma_j) = \frac{\partial \delta_i^{op}(\sigma_i, \sigma_j)}{\partial \sigma_i} d\sigma_i + \frac{\partial \delta_i^{op}(\sigma_i, \sigma_j)}{\partial \sigma_j} d\sigma_j \tag{30}$$

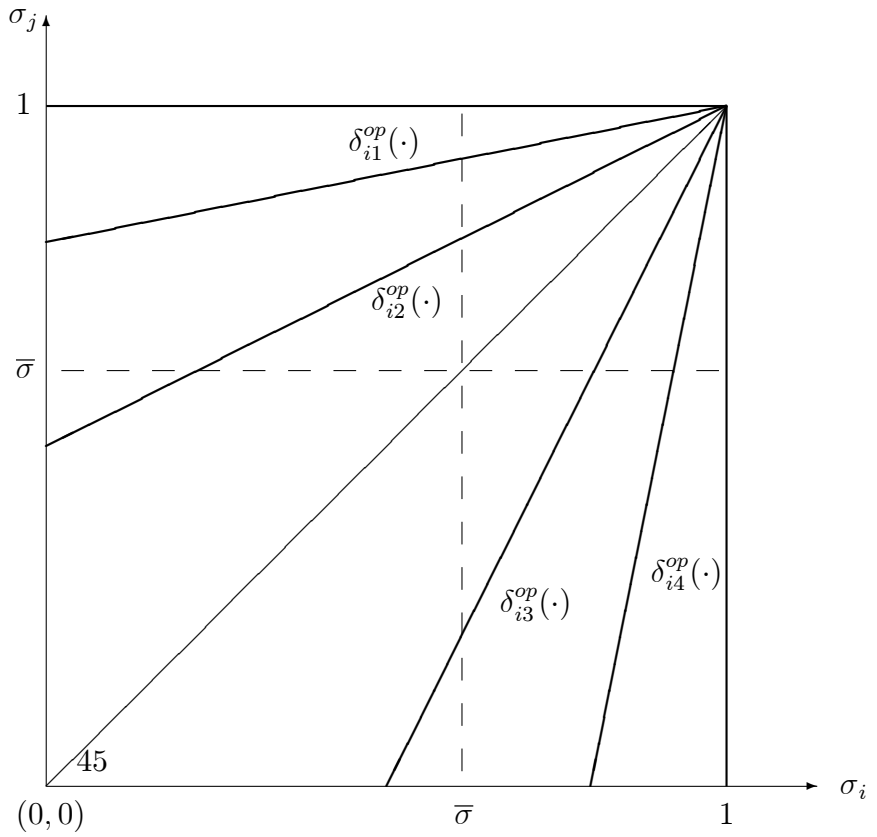
which may indeed take either sign. Then, imposing $d\delta_i^{op}(\sigma_i, \sigma_j) = 0$, obtain the marginal rate of substitution between σ_i and σ_j :

$$\frac{d\sigma_j}{d\sigma_i} = -\frac{\partial \delta_i^{op}(\sigma_i, \sigma_j)}{\partial \sigma_i} \cdot \frac{\partial \delta_i^{op}(\sigma_i, \sigma_j)}{\partial \sigma_j} = \frac{1-\sigma_j}{1-\sigma_i} > 0 \,\forall\, \sigma_i, \sigma_j \in [0,1), \tag{31}$$

illustrating the complementarity between a and m. Correspondingly, δ^* yields the map appearing in Figure 1, with

$$\delta_{i4}^{op}(\cdot) > \delta_{i3}^{op}(\cdot) > \delta_{i2}^{op}(\cdot) > \delta_{i1}^{op}(\cdot). \tag{32}$$

Figure 1: The critical discount factor of country i in the space (σ_i, σ_j)



The isoquants are upward sloping, as implied by the sign of (31). The 45-degree line represents the critical discount factor in the fully symmetric case previously examined. Hence, fixing $\sigma_j = \overline{\sigma}$ (as represented by the horizontal dashed line), an increase in σ_i makes it more difficult for country i to refrain from attacking. Conversely, supposing $\sigma_i = \overline{\sigma}$ and moving upward along the vertical dashed line one sees that an increase in σ_j for any given σ_i makes it increasingly easier for country i to satisfy the stability condition posed by

(27). The bottom line of this analysis is

Proposition 5 *Under optimal punishments, any departure from full symmetry decreases the stability of the international system by increasing one country's critical discount factor while decreasing the other.*

To appreciate the message conveyed by the above proposition, it suffices to observe that (28-29) also imply:

$$\frac{\partial \delta_i^{op}(\sigma_i, \sigma_j)}{\partial \sigma_i} > 0 \text{ and } \frac{\partial \delta_j^{op}(\sigma_i, \sigma_j)}{\partial \sigma_i} < 0 \,\forall\, \sigma_i, \sigma_j \in [0, 1) \tag{33}$$

so that any unilateral departure from the 45-degree line makes one country more aggressive and the other less so - acting as a puppy dog[11] - thereby creating a very risky situation in which pulling the trigger becomes a lot easier.

Now we may turn our attention to Friedman's (1971) grim trigger strategies. The Nash equilibrium payoff is

$$u_i^* = \pi - \frac{\left[2(1-\sigma_i)^2 - z(1-\sigma_j)\right]\pi^2 z}{4b} - C(\sigma_i) \tag{34}$$

and using it alongside with u_i^{PP} and u_i^{AP}, we can write the familiar stability condition:

$$\frac{u_i^{PP}}{1-\delta_i} \geq u_i^{AP} + \frac{\delta_i u_i^*}{1-\delta_i} \tag{35}$$

which delivers

$$\delta_i^g = \frac{z(1-\sigma_j)^2}{4(1-\sigma_i)^2} \tag{36}$$

as the critical threshold of the discount factor for country i (superscript g stands for *grim trigger strategies*). The above expression entails that (i) $\delta_i^g = \delta_i^{op}/2$ and (ii) δ_i^g generates a map of isoquants qualitatively analogous to Figure 1, and the same considerations as in Proposition 5.

[11] This label, identifying a player as a comparatively innocuous domestic pet, dates back to Fudenberg and Tirole (1984).

5 Concluding remarks

We have revisited the issue of the interplay between deterrence and defensive weapons in a repeated game with optimal one-shot punishments, to show that the sustainability of peace is unaffected by defensive shields if each country has one, while any uneven distribution of such weapons has indeed destabilising consequences. This provides a confirmation of the balance of power claim that symmetry is indeed conducive to stability. In our model, whet matters is the balance between defensive shields rather that offensive weapons, which may not be symmetrically distributed across countries.

A desirable extension of the foregoing analysis would consists in allowing for the presence of some realistic degree of attrition negatively affecting the performance of defensive shields during a war. An analogous problem may indeed affect each country's capability of replenishing its stock of nuclear weapons. An alternative and completely different approach to the unilateral *vs* generalised adoption of defensive weapons and the assessment of its effects on peace stability would consist in confining attention to the one-shot game and use a forward induction argument *à la* Kohlberg and Mertens (1986), as used in Lambertini (2013) to study the proliferation of nuclear weapons. These extensions are left for future research.

References

[1] Abreu, D.J. (1986), "Extremal Equilibria of Oligopolistic Supergames", *Journal of Economic Theory*, **39**, 191-225.

[2] Axelrod, R. (1981), "The Emergence of Cooperation among Egoists", *American Political Science Review*, **75**, 306-18.

[3] Brams, S.J. and D.M. Kilgour (1988), "Deterrence versus Defense: A Game-Theoretic Model of Star Wars", *International Studies Quarterly*, **32**, 3-28.

[4] Bundy, M., G.F. Kenna, R.S. McNamara and G. Smith (1984/1985), "The President's Choice: Star Wars or Arms Control", *Foreign Affairs*, **63**, 264-78.

[5] Carlsson, H. and E. van Damme (1993), "Global Games and Equilibrium Selection", *Econometrica*, **61** 989-1018.

[6] Chassang, S. and G. Padró I Miquel (2009), "Defensive Weapons and Defensive Alliances", *American Economic Review*, **99** (P&P), 282-86.

[7] Chassang, S. and G. Padró I Miquel (2010), "Conflict and Deterrence under Strategic Risk", *Quarterly Journal of Economics*, **125**, 1821-58.

[8] Danilovic, V. (2002), *When the Stakes are High: Deterrence and Conflict among Major Powers*, Ann Arbor, University of Michigan Press.

[9] Fearon, J.D. (1995), "Rationalist Explanations for War", *International Organization*, **49**, 379-414.

[10] Freedman, L. (2004), *Deterrence*, Cambridge, MA, Polity Press.

[11] Friedman, J.W. (1971), "A Non-Cooperative Equilibrium for Supergames", *Review of Economic Studies*, **28**, 1-12.

[12] Fudenberg, D. and J. Tirole (1984), "The Fat-Cat Effect, the Puppy-Dog Ploy, and the Lean and Hungry Look", *American Economic Review*, **74**, 361-66.

[13] Harsanyi, J. and R. Selten (1988), *A General Theory of Equilibrium Selection in Games, Cambridge*, MA, MIT Press.

[14] Klemperer, P. and M. Meyer (1986), "Price Competition vs Quantity Competition: The Role of Uncertainty", *RAND Journal of Economics*, **17**, 618-38.

[15] Kohlberg, E. and J.-F. Mertens (1986), "On the Strategic Stability of Equilibria", *Econometrica*, **54**, 1003-38.

[16] Lambertini, L. (2013), "Joining the Nuke Club: A Forward Induction Approach", *Defence and Peace Economics*, **24**, 15-21.

[17] Lebovic, J.H. (2002), "The Law of Small Numbers: Deterrence and National Missile Defense", *Journal of Conflict Resolution*, **46**, 455-83.

[18] Lindsay, J.M. and M.E. O'Hanlon (2001), *Defending America: The Case for Limited National Missile Defense*, Washington, DC, Brookings Institution Press.

[19] Miller, S.E. (2001), "The Flawed Case for Missile Defense", *Survival*, **43**, 95-109.

[20] Mills, M.J., O.B. Toon, R.P. Turco, D.E. Kinnison and R.R. Garcia (2008), "Massive Global Ozone Loss Predicted following Regional Nuclear Conflict", *Proceedings of the National Academy of Sciences*, **105**, 5307-12.

[21] Powell, R. (1990), *Nuclear Deterrence Theory: The Search for Credibility*, Cambridge, Cambridge University Press.

[22] Powell, R. (1996), "Stability and the Distribution of Power", *World Politics*, **48**, 239-67.

[23] Powell, R. (1999), *In the Shadow of Power: States and Strategies in International Politics*, Princeton, NJ, Princeton University Press.

[24] Powell, R. (2003), "Nuclear Deterrence Theory, Nuclear Proliferation, and National Missile Defense", *International Security*, **27**, 86-118.

[25] Powell, R. (2006), "War as a Commitment Problem", *International Organization*, **60**, 169-203.

[26] Quackenbush, S.L. (2006), "National Missile Defense and Deterrence", *Political Research Quarterly*, **59**, 533-41.

[27] Quackenbush, S.L. and A.C. Drury (2011), "National Missile Defense and (Dis)satisfaction", *Journal of Peace Research*, **48**, 469-80.

[28] Rhodes, R. (1995), *Dark Sun. The Making of the Hydrogen Bomb*, New York, Simon and Schuster.

[29] Schelling, T. (1960), *The Strategy of Conflict*, Cambridge, MA, Harvard University Press.

[30] Wilkening, D.A. (2000), "Ballistic-Missile Defense and Strategic Stability", Adelphi Paper no. 334, London, International Institute for Strategic Studies.

[31] Zagare, F.C. (1996), "Classical Deterrence Theory: A Critical Assessment", *International Interactions*, **21**, 365-87.

[32] Zagare, F.C. (2004), "Reconciling Rationality with Deterrence: A Reexamination of the Logical Foundations of Deterrence Theory", *Journal of Theoretical Politics*, **16**, 107-41.

[33] Zagare, F.C. and D.M. Kilgour (1993), "Asymmetric Deterrence", *International Studies Quarterly*, **37**, 1-27.

[34] Zagare, F.C. and D.M. Kilgour (2000), *Perfect Deterrence*, Cambridge, Cambridge University Press.