# Anti-terrorism policies and the risk of provoking

Franz Dietrich

May 2008 (minor revisions later)

forthcoming in *Journal of Theoretical Politics*

**Abstract.** Tough anti-terrorism policies are often defended by focusing on a fixed minority of the population who prefer violent outcomes, and arguing that toughness reduces the risk of terrorism from this group. This reasoning implicitly assumes that tough policies do not increase the group of 'potential terrorists', i.e., of people with violent preferences. Preferences and their level of violence are treated as stable, exogenously fixed features. To avoid this unrealistic assumption, I formulate a model in which policies can 'brutalise' or 'appease' someone's personality, i.e., his preferences. This follows the endogenous preferences approach, popular elsewhere in political science and economics. I formally decompose the effect of toughness into a (desirable) deterrence effect and an (undesirable) provocation effect. Whether toughness is overall efficient depends on which effect overweighs. I show that neglecting provocation typically leads to toughness exaggeration. This suggests that some tough anti-terrorism policies observable in the present and past can be explained by a neglect of provocation.

**Keywords.** Terrorism, endogenous preferences, reciprocity, dynamic inconsistency

## 1 Introduction

The public debate on terrorism policies is dominated by two goals: protecting society against *existing* terrorists, and preventing the emergence of *new* terrorists. The first goal initially suggests a tough policy of fighting terrorists and deterring them by severe punishments. But, while toughness might often be successful with respect to the first goal, it can undermine the second goal by provoking individuals who were previously peaceful. The heart of the disagreement over the correct anti-terrorism policy, in politics and society, is that proponents of toughness usually refer to the first goal, whereas critics of

toughness usually refer to the second goal. Awareness for the other goal is often lacking in each camp. While the public debate stays informal, it is important to formally analyse the two-edged effect of toughness. This has to happen in a single model since we need to understand the interaction and trade-off between the two effects. But does there exist any compelling model of two phenomena as different as deterrence and provocation? Many increasingly refined game-theoretic treatments, for instance, are quite successful in capturing deterrence effects, but (with few exceptions) neglect provocation: they assume that the risk of terrorism comes from a fixed group of individuals with stable violent preferences and hence optimize the policy with respect to these individuals only, neglecting potential provocation of other individuals who are not included as players in the model.

This theoretical paper proposes a new model that captures both (desirable) deterrence and (undesirable) provocation by anti-terrorism policies. The model construes provocation as a form of preference change. It considers a large group of individuals (possibly the entire world population and notably all potentially provocable individuals) and treats as endogenous whether and how much someone likes violent outcomes: These preferences react to the anti-terrorism policy, reflecting that policies may create or reduce hate and other human feelings, particularly if these feelings are directed towards the policy makers themselves or the countries or cultures associated with them. Some policies might 'brutalise' preferences, others might 'appease' them.

In this model, policies can deter and provoke, and these two opposed policy effects can be quantified and weighed against each other. This allows one to conceptualise and compare the two opposed aspects of toughness (deterrence and provocation) and thereby places the popular disagreement over the effectiveness of toughness on formal grounds. To avoid distractions from our focus (on the two-edged effect of toughness), the model is kept abstract and simple on all other dimensions, emphasising mathematical generality while sacrificing concreteness and specificity. I avoid any *ad hoc* technical restrictions (e.g., to utility functions from a particular parametric class) or interpretational restrictions (e.g., to particular kinds of terrorism or toughness).

I prove that, under regularity conditions, the policy's effect on aggregate terrorism is the sum of its deterrence effect and its provocation effect, and, further, that under plausible (but not universal) conditions the deterrence effect is negative (i.e., terrorism-reducing) and the provocation effect positive (i.e., terrorism-increasing). So, the policy maker effectively faces a trade-off between deterrence and provocation. I also compare provocation-aware with provocation-neglecting policies and prove that provocation-neglect quite gener-

ally leads to toughness exaggeration.

The paper's analysis invites comparisons with contemporary and past anti-terrorism policies, including the question of how much these policies have provoked (in the paper's technical sense) and whether their choice was driven by provocation-neglect. Such comparisons with reality are left to the reader's imagination and to empirical follow-up work. Being non-empirical, the present paper refrains from concrete empirical claims.

The paper is organised into an informal part (Section 2) that introduces many key ideas, and a formal part (Section 3) that provides the mathematical foundations. In Appendix A, an application is worked out. In Appendix B, all proofs are given.

Although provocation by anti-terrorism policies is discussed in political science and sociology (and by people on the streets), formal rational-choice-based models usually ignore it (some exceptions are mentioned below). In focussing on provocation and modelling it as a form of preference/taste change, I take the approach of the literature on endogenous preferences. The instability of tastes and their endogenous determination by environmental factors such as governmental policies or institutions is increasingly recognised and modelled in economics (e.g., Polak 1976, Hansson 1995, Becker 1996, Bowles 1998, Rabin 1998, Dekel et al. 2007, Dietrich 2012). It is important to incorporate this approach into terrorism modelling, because dispositions towards terrorism seem particularly unstable and environment-sensitive in that they typically reflect complex mental states rather than basic biological attributes or needs.

Provocation as modelled here (that is: a policy-caused development of preferences for violent outcomes) can be interpreted *either* as a rational taste change, derived from stable extended preferences over extended alternatives that contain the policy as a taste parameter, *or* as a dynamically inconsistent taste change. The first interpretation allows one to explain provocation by a stable preference to reciprocate, i.e. to harm tough policy makers (countries, cultures etc.) and to be mild to soft ones. The rational-choice foundations of reciprocal feelings are understood increasingly well (e.g., Rabin 1993, Fehr and Gächter 1998, Bolton and Ockenfels 2000, Sethi and Somanathan 2001, 2003, Dufwenberg Kirchsteiger 2004 and Falk and Fischbacher 2006). The second, no less important interpretation of provocation, namely dynamic inconsistency, stresses the effect that anti-terrorism policies can have on someone's personality and psychological state. This results in a change in *fundamental* preferences (e.g., Strotz 1955-56, Hammond 1976, O'Donoghue and Rabin 1999, Bénabou and Pycia 2002). One might defend such an interpretation of provocation against one in terms of stable reciprocal preferences by arguing that

terrorists were usually not born with the (conditional) preference to perform terrorist attacks in future conditional on such and such future circumstances; radical preferences are usually created over time by circumstances (e.g., by a war) rather than being always present as conditional preferences, so the argument.

By standing in the tradition of the endogenous preference literature, this paper is less related to existing work on terrorism prevention in political economy. One branch of this field is empirical and investigates potential causes of terrorism; e.g., Eubank and Weinberg (1994), Silke (1994), and Dumas (2002). A more theoretic (often *game*-theoretic) branch assesses the efficiency of various anti-terrorism measures by accounting for costs and/or various strategic incentives of terrorists and governments. Some important contributions are Brams and Kilgour (1985, 1987), Cioffi-Revilla (1985, 1998), Lichbach (1987), Zagare and Kilgour (2000) (with the perhaps most prominent theory of interstate deterrence), Frey and Luechinger (2002), Frey (2004), Enders and Sandler (2006), Goodin (2006). Part of this literature emphasizes limitations of deterrence and toughness, yet not on grounds of resulting provocation (in our sense), but for instance on grounds of non-credibility of certain threats or on grounds of costs. A few game-theoretic contributions do however account for provocation in their own ways; in particular, Brams and Kilgour (1988), Rosendorff and Sandler (2004) and Bueno de Mesquita (2005) study models in which harsh governmental policies may increase the motivation of, or the support for, or the mobilisation of terrorists. Overall, the empirical and theoretical literature provides several important insights not reviewed here; future research might combine them with the insights on provocation effects developed here.

# 2 Informal analysis

Throughout the paper we consider a policy maker (e.g. a national government or international organisation) in charge of choosing and implementing an anti-terrorism policy. The term 'anti-terrorism policy' is understood in a broad sense (made precise in Sections 2.2 and 2.3), possibly including measures as different as the creation of social institutions, changes in the education system, military interventions, police presence, criminal legislation and jurisdiction, declarations and speeches by politicians, diplomatic relations, embargoes, and so on.

The terrorism threat comes from the members of a (finite non-empty) set of individuals $N$, called the *population*. Crucially, $N$ contains not just individuals currently engaged in terrorist activities (arguably a frequent mistake) but also all potential ones. This speaks for a large definition of $N$: it might include all

humans on earth, or some group defined geographically, ethnically, religiously, or else. The term 'individual' always refers to members of $N$.

In response to the policy, each individual engages in some behaviour, which can be more or less violent; he[1] might exercise no violence at all, or perform small offences, or major terrorist attacks, and so on. I assume that an individual cares about two consequences of his behaviour: (i) a level of damage created, represented by a real number $x \geq 0$, and (ii) a level of punishment received, represented by another real number $y \geq 0$. The term 'punishment' is used very broadly: it stands for *any* personal disadvantage ('cost') incurred, such as having to hide from authorities (before or after damage creation), having to prepare the attack, coming to prison afterwards, and so on. Usually, most individuals behave so as to obtain the no-damage-no-punishment outcome $(x, y) = (0, 0)$. I use the term 'terrorist' resp. 'non-terrorist' in a technical sense to denote someone who causes positive damage $x > 0$ resp. zero damage $x = 0$ (without intending any further connotations that this sensitive terminology may have in normal language). Although I say throughout that someone 'chooses' his damage-punishment pair $(x, y)$, this pair is in fact the outcome not just of own behaviour but also of the policy: damage $x$ could depend on the level of protection of targets, and punishment $y$ on criminal legislation. (In practice, the outcome $(x, y)$ often also depends on chance, since the exact level of damage $x$ created by behaviour under a policy is often subject to uncertainties, as is the level of punishment $y$, which might depend on whether the terrorist is captured and how lucky he is in his trial. Our framework can capture 'chance' if one re-interprets $(x, y)$ as an *expected*-damage-*expected*-punishment pair.[2])

## 2.1 Peaceful vs. brutal preferences

Suppose a given policy is in action. As usual, individuals are preference-maximisers. Accordingly, let each individual $i$ in the population $N$ hold some

[1] Throughout I use masculine pronouns, without intended gender restriction.

[2] To make our analysis compatible with this re-interpretation, one would need two technical assumptions: (i) an individual's behaviour leads to a (discrete or continuous) probability distribution (lottery) over damage-punishment pairs with a finite expectation in each coordinate; (ii) the individual ranks such lotteries based on their expectation pairs. Condition (ii) is restrictive, since it implies risk-neutrality: the agent must for instance be indifferent between achieving $(x, y) = (2, 2)$ *for sure* and achieving $(2, 0)$ or $(2, 4)$ with equal probabilities. Condition (ii) holds if the agent maximizes the expectation of a linear utility function $u(x, y) = ax + by$ (for constants $a, b \in \mathbb{R}$), but is violated for non-linear utility functions such as that given by $u(x, y) = \sqrt{x} + \sqrt{y}$ and that given by $u(x, y) = xy$.

preference order[3] $\succeq = \succeq_i$ on the set

$$\mathbb{R}^2_+ = [0,\infty) \times [0,\infty) = \{(x,y) : x \geq 0, y \geq 0\} \text{ (\textit{damage-punishment quadrant})}$$

of damage-punishment pairs $(x, y)$; the associated relations of strict preference $\succ$ and indifference $\sim$ are defined as usual.[4] As illustrated in Figure 1, some individuals might hold peaceful preferences, others brutal ones. Formally, I call a preference order $\succeq$ (on the damage-punishment quadrant $\mathbb{R}^2_+$)

- *peaceful* if less damage is preferred to more ceteris paribus, i.e. if $(x, y) \succ (x', y)$ whenever $x < x'$, and *brutal* otherwise.

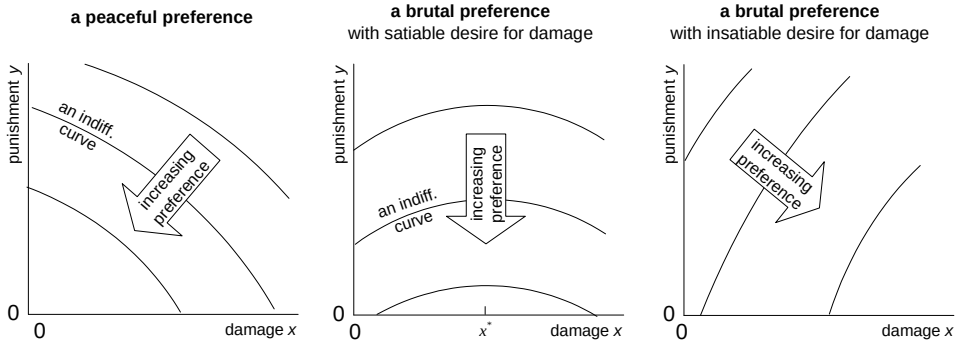


Figure 1: A peaceful and two brutal preferences (all three punishment-averse)

Someone's preference – whether peaceful or brutal – is usually punishment-averse, where I call a preference order $\succeq$ (on the damage-punishment quadrant $\mathbb{R}^2_+$)

- *punishment-averse* if less punishment is preferred to more ceteris paribus, i.e. if $(x, y) \succ (x, y')$ whenever $y < y'$.

Given punishment-aversion (which I assume throughout the informal discussion), the difference between peaceful and brutal preferences shows in the slope of indifference curves: peaceful preferences have negatively sloped indifference curves (first plot in Figure 1), whereas brutal preferences have positively sloped indifference curves at least somewhere on the quadrant $\mathbb{R}^2_+$ (second and third plot in Figure 1). By anti-clockwise rotating the indifference curves, preference becomes *more brutal*, where I have in a natural way (partially) ordered the preference orders on $\mathbb{R}^2_+$ in terms of brutality:

- $\succeq$ is *at least as brutal* (or *at most peaceful*) as $\succeq'$ if any preference for higher damage that holds under $\succeq'$ also holds under $\succeq$, i.e. if $(x, y) \succeq' (x', y')$ and $x > x'$ imply $(x, y) \succeq (x', y')$.

[3] Throughout, 'preference order' refers to a transitive and complete binary relation.

[4] For all $(x, y), (x', y') \in \mathbb{R}^2_+$, we have $(x, y) \succ (x', y') \Leftrightarrow [(x, y) \succeq (x', y')$ and not $(x', y') \succeq (x, y)]$, and $(x, y) \sim (x', y') \Leftrightarrow [(x, y) \succeq (x', y')$ and $(x', y') \succeq (x, y)]$.

In this sense, the preference on the right of Figure 1 is more brutal than that in the middle, which is more brutal than that on the left. A radical form of brutal preferences $\succeq$ are ones with

- *insatiable damage desire*, i.e. $(x, y) \succ (x', y)$ whenever $x > x'$,

in which case indifference curves are positively sloped on the entire quadrant $\mathbb{R}^2_+$ (third plot in Figure 1). (One might speculate whether some suicide bombers have insatiable damage desire.) The more moderate forms of brutal preferences are ones with a strictly positive but finite optimal damage level $x^*$ (which might depend on punishment $y$), where preference typically decreases as damage $x$ moves away from the optimum in either direction holding $y$ fixed (see second plot in Figure 1). For instance, someone might desire to destroy a building, but preferably without killing humans.

Each anti-terrorism policy results in some (non-empty) set $\mathbf{F} \subseteq \mathbb{R}^2_+$ of *feasible* damage-punishment pairs from which individuals have to 'choose'. Figure 2 and later figures take the feasible set $\mathbf{F}$ to be 'thin' and linear and to render every damage level $x \geq 0$ feasible (nothing of which is essential[5]), with the plausible feature that punishment increases with damage and that no-damage-no-punishment $(0, 0)$ is feasible. Figure 2 shows how three types of individuals

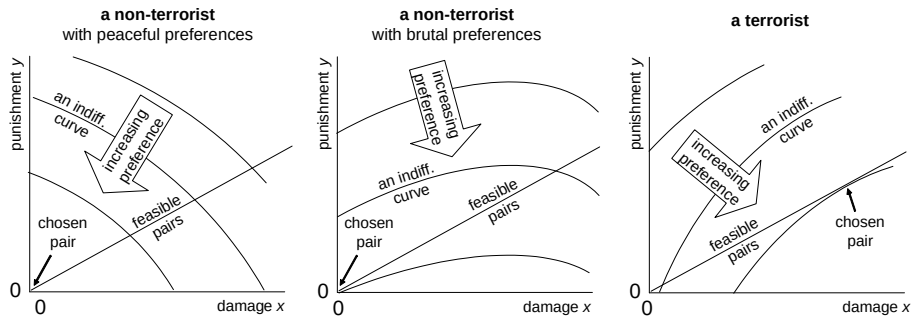


Figure 2: Behaviour of one peaceful and two brutal types under the policy

behave under the policy (all maximising preference): the peaceful type on the left creates no damage (is not a terrorist), the strongly brutal type on the right creates positive damage (is a terrorist), and the moderately brutal type in the middle creates no damage (but would have been a terrorist under only slightly more brutal preferences).

[5]Feasible sets are discussed in full generality in Section 3.1.

## 2.2 Cause-related and symptom-related policy measures

A policy measure can qualify as part of the *anti-terrorism* policy if, through whatever means, it affects the damage level created by individuals. Someone's damage level $x$ is determined by two factors: (i) his preference order $\succeq$ on the damage-punishment quadrant $\mathbb{R}^2_+$ and (ii) the feasible set $\mathbf{F} \subseteq \mathbb{R}^2_+$ from which he chooses. This naturally leads me to distinguish between two sorts of anti-terrorism policy measures, to be labelled 'cause-related' or 'symptom-related' (without the derogative connotation that the term 'symptom-related' sometimes has in natural language):

- A *cause-related* or *appeasement* measure aims at changing the preferences of population members, not the constraints $\mathbf{F} \subseteq \mathbb{R}^2_+$ under which they can create terrorism. The goal is that brutal preferences become peaceful (as on the right in Figure 3) or at least less brutal (as on the left in Figure 3). The question as to which measures succeed in appeasing preferences is bound to be context-dependent and controversial: Should one focus on improving education? or on raising the standard of living? or on reducing polarisation? or on inducing sympathy (say using advertisements) with those persons, cultures or institutions against which terrorism might be directed? or on reducing media attention from terrorism to render the latter less desirable, as Bruno Frey (2004) advocates? and so on. There is mixed evidence on how such measures affect a different but related variable, terrorism itself (rather than preferences); for instance, Krueger and Malecková (2003) cast doubts on the efficiency of raising education and lowering poverty.

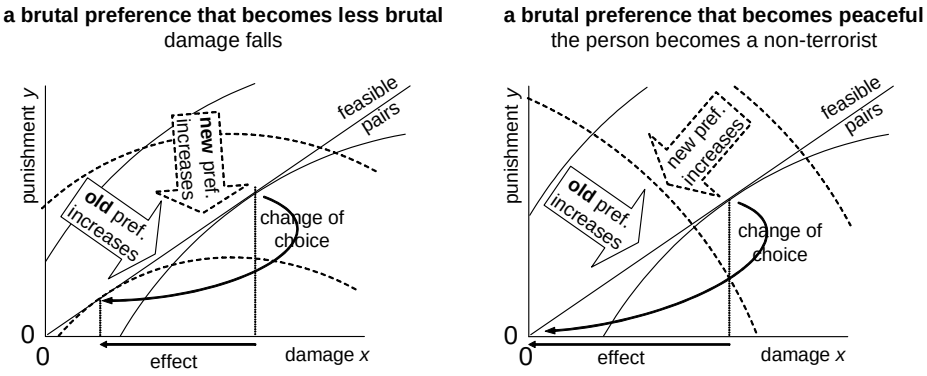


Figure 3: A succesful cause-related policy measure (new preference dashed)

- A *symptom-related* or *deterrence* measure aims to reduce terrorism by changing the constraints $\mathbf{F} \subseteq \mathbb{R}^2_+$ under which terrorists operate (possibly with brutalising side effects on people's preferences, as analysed in the next

subsection). Roughly, such measures render the feasible set $\mathbf{F}$ 'steeper', and perhaps render some damage levels $x$ infeasible.[6] Such measures can be *defensive* or *aggressive*. Defensive measures change the difficulty of creating damages $x$, for instance by erecting weapons embargoes, protecting buildings, supervising public places, enforcing transparency in bank transfers, or decentralising society to make it less vulnerable as Bruno Frey (2004) proposes. Aggressive measures change the kind or extent of punishment $y$, for instance by severe legislation, a worldwide search for terrorists, or a war. While there may be overlaps, the difference between defensive and aggressive deterrence can be formalised (see Section 4). A related distinction in the literature is that between 'proactive' and 'defensive' measures (e.g., Rosendorff and Sandler, 2004).

## 2.3 Symptom-related measures and the problem of provocation (side) effects

The rest of the paper (except Appendix A) analyses *symptom*-related policy measures as just defined. As Figure 4 illustrates, additional toughness does not lead to more terrorism if people's preferences are guaranteed to be non-provocable, i.e. policy-invariant. However, as Figure 5 illustrates, symptom-

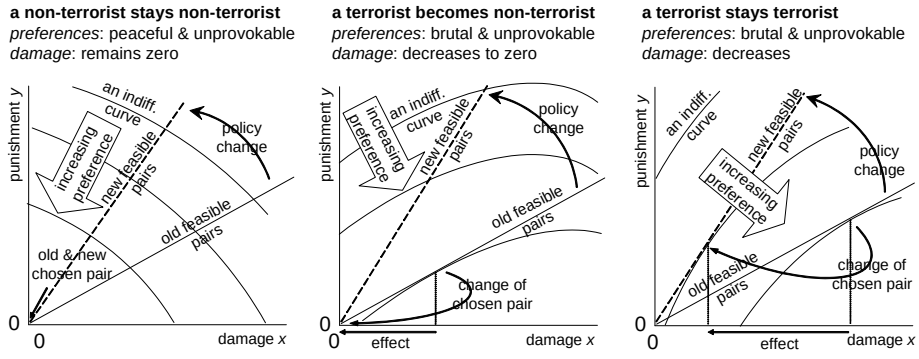


Figure 4: Effect of a toughness raise on three types of non-provocable individuals

related policies often affect not just the set $\mathbf{F} \subseteq \mathbb{R}^2_+$ of feasible damage-punishment pairs but (as side effects) also some individuals' preferences. The same individual who in the status quo still holds peaceful preferences $\succeq$ on $\mathbb{R}^2_+$ (hence is not a terrorist) might develop brutal preferences $\succeq'$ on $\mathbb{R}^2_+$ under a new tougher policy. It is psychologically plausible and empirically observable that some persons'

[6]Damage level $x \geq 0$ is feasible if $(x, y) \in \mathbf{F}$ for some punishment $y \geq 0$. Some policies (e.g. weapons embargoes) render high damage levels infeasible; see Section 3.1.

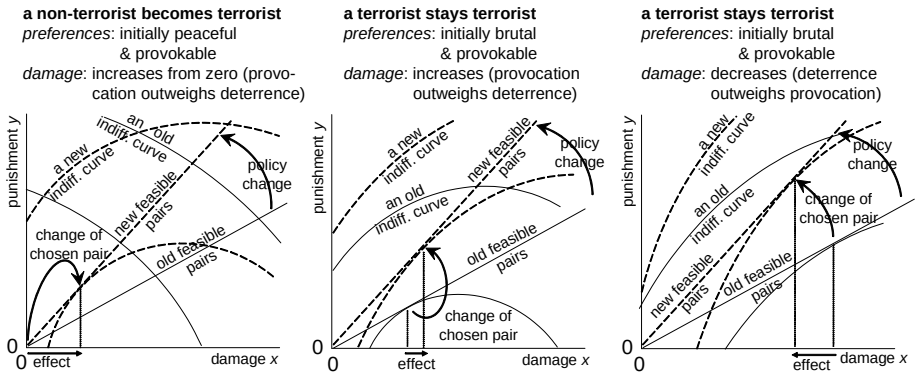


Figure 5: Effect of a toughness raise on three types of provocable individuals

preference for or against exercising violence is not stable but reacts to the environment: some environments appease, others brutalise tastes and desires. The anti-terrorism policy may form part of the environment that shapes someone's personality and preferences. A policy may let someone develop hate feelings and a preference for creating damage, either in order to hurt the policy-makers themselves (if the hate feelings focus on them) or without a specific target (if the hate feelings are more diffuse). Such provocation is given two interpretations in Section 2.4, namely in terms of either dynamic inconsistency or extended preferences over damage-punishment-policy triples (the latter interpretation offering two perfectly rational explanations of provocation: reciprocity and taste acquisition).

This said, an individual's preference order on the damage-punishment quadrant $\mathbb{R}^2_+$ should be indexed by the policy $t$, say $\succeq_t$ with $t$ ranging over a set $T$ of relevant (symptom-related) policies among which the status quo policy $\bar{t}$. Here, $\succeq_t$ represents the individual's (more or less peaceful) dispositions under (the impression of) policy $t$. I call an individual, or the family of his policy-indexed preference orders $(\succeq_t)_{t\in T}$ on $\mathbb{R}^2_+$,

- *unprovocable* if preference $\succeq_t$ is the same for each policy $t \in T$;
- *provocable* otherwise.[7]

This definition of provocability leaves open the direction of the preference reaction. In principle, provocation could even take the inverse form that tough policies appease preferences (such as when someone becomes unable to touch a knife after the traumatic experience of a war); this psychological reaction, later

[7]This terminology makes sense since $T$ consists of *symptom*-related policies. But if policies may differ also (or only) in cause-related measures (e.g. in the education system), the more general term '(un)changeable' is better than '(un)provocable'. Changing (appeasing) preferences is the whole point of cause-related anti-terrorism policies.

referred to as 'inverse provocation', is possible but seems less frequent.

If one were to decompose someone's preferences into a damage-related utility and a punishment-related utility, say, through an additively separable utility model $u_t(x, y) = v_t(x) + w_t(y)$ as in Appendix A, then one is led to ask: does provocability come rather from damage utility $v_t(x)$ being policy-sensitive, or rather from punishment utility $w_t(y)$ being policy-sensitive? Certainly, utility derived from terrorism $v_t(x)$ seems more likely to be policy-sensitive: while the pain from a fixed punishment level $y$ seems policy-invariant, the pleasure of creating a fixed damage $x$ highly depends on how much the person dislikes the policy makers (cultures, etc.) he hurts, which may be policy-sensitive.

How does a policy toughening affect terrorism? As illustrated in Figure 4, the behaviour of *unprovocable* individuals is affected in a desirable way: non-terrorists stay non-terrorists, and terrorists reduce damage, possibly becoming non-terrorists. Such deterrence without provocation is extensively studied in the crime and terrorism literature, using different models. For provocable types, the picture changes and becomes less uniform. Damage increases for types where provocation outweighs deterrence (first two plots of Figure 5), but decreases for types where deterrence outweighs provocation (third plot in Figure 5). In Figure

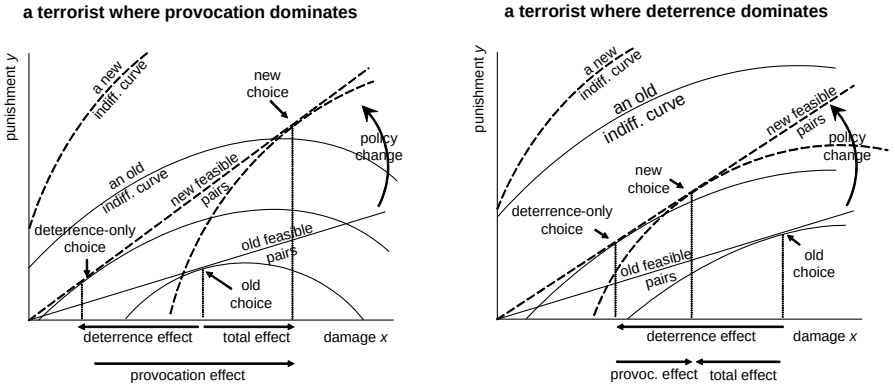


Figure 6: Deterrence effect and porvocation effect

6, I decompose the total effect of the toughness rise on a terrorist's damage into

- a (usually negative) *deterrence effect*, representing the damage change if, hypothetically, preferences were to remain constant, and
- a (usually positive) *provocation effect*, caused by preference change.

There is no general rule as to which of these two competing effects is stronger; for the type on the left (right) in Figure 6, provocation (deterrence) is stronger. The same decomposition also works for non-terrorists: the deterrence effect is then zero (less than 'no damage' doesn't exist) but the provocation effect might

be positive, turning the person into a terrorist.

I shall be formal in Section 3 about the decomposition into deterrence and provocation. But qualitatively, what can we learn already now? If the population is approximately homogeneous, the policy maker can reduce terrorism by adjusting the policy to the predominant type; for instance, high toughness is efficient against a population dominated by types that are unprovocable (see Figure 4) or little provocable (see right plots in Figures 5 and 6). Often though, the population is significantly heterogeneous and contains many types (preferences), some more provocable than others, some peaceful and others brutal in the status quo. Then the policy maker faces the difficult task of finding the right compromise given the type distribution. As a rule of thumb, optimal toughness is decreasing as a function of the level of provocability and increasing as a function of the level of (status quo) brutality in the type distribution. More precisely, a policy shift from the status quo $\bar{t}$ to some new policy $t \in T$ minimises *sum-total* terrorism, as given by the sum $\mathbf{x} = \sum_{i \in N} x_i$ of damage levels $x_i$ across individuals $i \in N$, if and only if it minimises the policy's *aggregate* effect on damage (i.e. the change of $\mathbf{x}$ from the status quo). This effect can be decomposed into the sum $\mathbf{DE}(t) + \mathbf{PE}(t)$ of the aggregate deterrence effect $\mathbf{DE}(t)$ and the aggregate provocation effect $\mathbf{PE}(t)$.[8] $\mathbf{DE}(t)$ and $\mathbf{PE}(t)$ measure how much policy $t$ affects terrorism through deterrence and provocation, respectively. In practice, the policy maker faces two distinct challenges:

*Finding the right level of toughness.* Suppose the policy maker decides on a *single* policy parameter (e.g. the size of a military intervention), so that we can identify policies with toughness levels $t$ chosen from a one-dimensional policy space $T \subseteq \mathbb{R}$. As Figure 7 illustrates, in trying to minimise the function

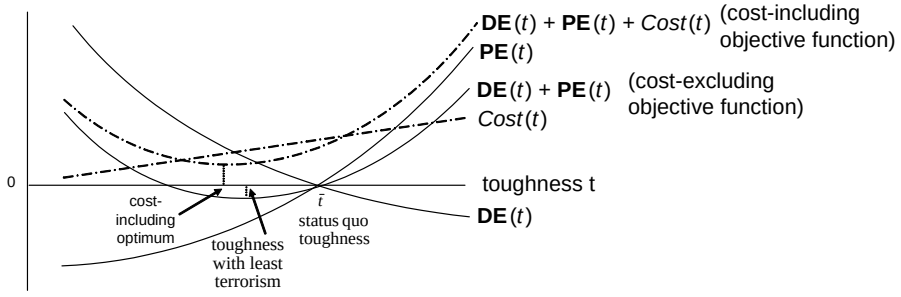


Figure 7: An example in which reducing toughness reduces terrorism

[8] That is, $\mathbf{DE}(t) = \sum_{i \in N} DE_i(t)$ where $DE_i(t)$ is individual $i$'s deterrence effect, defined as his change of damage holding his preferences fixed, i.e. neglecting provocation. Similarly, $\mathbf{PE}(t)$ is the sum of the individual provocation effects $PE_i(t)$, $i \in N$.

$\mathbf{DE}(t)+\mathbf{PE}(t)$, the policy maker faces a standard one-dimensional trade-off because $\mathbf{DE}(t)$ is decreasing but $\mathbf{PE}(t)$ increasing in toughness $t$ (for details, see Section 3). A minimum of $\mathbf{DE}(t)+\mathbf{PE}(t)$ defines an optimum on the trade-off between deterrence and provocation. Overshooting toughness increases terrorism by provoking too much, and undershooting toughness increases terrorism by deterring too little. As also illustrated in Figure 7, a refined objective function might be $\mathbf{DE}(t)+\mathbf{PE}(t)+Cost(t)$, where $Cost(t)$ represents the (suitably scaled) cost of toughness $t$ (such as financial costs, loss of human lives, and loss of life quality through state supervision). As $Cost(t)$ typically increases in toughness $t$, the optimum is typically reached at lower toughness than under the cost-neglecting objective function $\mathbf{DE}(t)+\mathbf{PE}(t)$. In short, toughness $t$ should be the lower, the steeper the $\mathbf{PE}(t)$ curve is (more provocability), the flatter the $\mathbf{DE}(t)$ curve is (less deterrability), and the steeper the cost curve $Cost(t)$ is (more costly toughness).

*Finding the right kind of toughness.* Suppose now that a policy $t$ is given by many policy parameters: criminal legislation, weapons embargoes, police presence, military interventions, and so on. This may be represented by a multi-dimensional policy space $T$.[9] Interestingly, the same level of deterrence – i.e. the same set $\mathbf{F}$ of feasible damage-punishment pairs, hence the same deterrence effect $\mathbf{DE}(t)$ – is often achievable by several policies $t \in T$ that differ in their dimension-specific toughness levels and thereby provoke in different ways and to different overall extents $\mathbf{PE}(t)$: some of these policies may lead to peaceful preferences on $\mathbb{R}^2_+$ for most individuals, others to many brutal preferences. On which dimensions should the policy be tough, on which mild? Our model recommends a policy that achieve its overall level of deterrence in the least provoking way, which is implemented by allocating toughness to dimensions where deterrence comes with little provocation. The reason is that different policies $t$ with same aggregate deterrence effect $\mathbf{DE}(t)$ can be compared based just on their aggregate provocation effect $\mathbf{PE}(t)$ (possibly plus policy costs $Cost(t)$). For instance, if introducing a weapons embargo leaves most individuals' preferences either totally unchanged (as in Figure 3) or brutalises them just slightly in the sense that deterrence overweighs provocation (as in the right plots of Figures 5 and 6), then *this* toughness raise is desirable, and preferable to other ones that deter equally but provoke more.[10] Whether *overall* deterrence should be high (i.e. $\mathbf{F}$ should be 'steep') is context-dependent.

[9] That is, $T \subseteq \mathbb{R}^k$, where a policy is seen as a vector $t = (t_1, ..., t_k)$, and $t_1$ is the level of toughness of 1st kind, $t_2$ that of 2nd kind, and so on.

[10] The embargo does indeed deter: the feasible set $\mathbf{F} \subseteq \mathbb{R}^2_+$ gets 'steeper' because damage creation gets harder (if feasible at all) and more criminal (so more highly punished).

## 2.4 Provocable preferences: a case of rationality or of dynamic inconsistency?

What can make someone's preference over damage-punishment pairs react to the policy (e.g. to a war)? I deliberately leave the paper's analysis compatible with two classical economic interpretations of preference change: *dynamic inconsistency*, and what I call *rational taste change* in deference to Gary Becker's terminology.

*Rational taste change: reciprocity and acquired tastes.* Under this interpretation, an individual's policy-dependent preference orders $\succeq_t$, $t \in T$, (on the damage-punishment quadrant $\mathbb{R}^2_+$) are derived from a single stable *extended* preference order $\tilde{\succeq}$ over the set $\mathbb{R}^2_+ \times T$ of damage-punishment-policy triples $(x, y, t)$, in which $t$ plays the role of a *taste parameter*: for each $t \in T$, $(x, y) \succeq_t (x', y')$ then simply means that $(x, y, t) \tilde{\succeq} (x', y', t)$, i.e. that the individual prefers having $(x, y)$ *with policy* $t$ to having $(x', y')$ *with policy* $t$ (just as someone might prefer white to red wine *with desert*, though perhaps not with cheese). Such extended preferences can rationalise provocation, in at least two ways. First, there may be a desire to reciprocate, i.e. to be violent to tough policy makers (foreigners, etc.) and peaceful to mild ones. Second, the ability to enjoy terrorism may be *acquired*, namely through experiencing the policy, e.g. a war (just as a consumer à la Becker acquires the ability to enjoy a good by building up a stock of social and personal capital); note that taste acquisition does not imply dynamic inconsistency (just as the Becker consumer is not dynamically inconsistent: he anticipates his future abilities). The indifference curves of $\succeq_t$ (plotted in our figures) are derived from the higher-dimensional indifference sets of $\tilde{\succeq}$ (in the space $\mathbb{R}^2_+ \times T$) by fixing the 'third coordinate' $t$, i.e. by intersecting with the subspace $\mathbb{R}^2_+ \times \{t\}$.

*Dynamic inconsistency.* A dynamically inconsistent agent disapproves of his own future preference: he undergoes a personality change under the impression of the changing environment (policy), and his preference change is not representable by stable extended preferences (or stable intertemporal preferences over complete event streams). To illustrate, suppose the policy changes from the status quo 'mild' ($\bar{t}$) to 'tough' ($t$). How would someone whose preference changes from 'peaceful' ($\succeq_{\bar{t}}$) to 'brutal' ($\succeq_t$) describe himself before the change?

- In the earlier case of rational taste change, he might say: "I want and will always want to harm tough foreigners and to be kind to mild ones, and so I am currently not a terrorist but intend to become one whenever the foreigners becomes tough."

- In the case of dynamic inconsistency, he might say: "I am currently a pacifist who does not want to harm any mild or tough foreigners, and I wish I could prevent that, once foreigners become tough, my personality changes and I develop a desire to harm them."

More formally, the person's *present* preference about his *future* damage-punishment under future policy $t$ are given by:

- present taste $\succeq_{\bar{t}}$ in the case of dynamic inconsistency;
- future taste $\succeq_t$ in the case of rational taste change;
- some combination of $\succeq_{\bar{t}}$ and $\succeq_t$ in mixed cases.

Empirically, the difference between these kinds of preference change is revealed in commitment behaviour.[11] In economics, dynamic inconsistency is often associated with individuals whose mental state is subjected to chocks or influences, either of an external kind (brutal friends, war) or an internal kind (Alzheimer, puberty). In this sense, provocation seems a natural candidate for dynamic inconsistency.

Which explanation of provocation, then, is appropriate? Answers are likely to be both context-dependent and controversial; the reader might choose his or her preferred explanation. It might even be that (within the same application) some individuals undergo a rational taste change and others a dynamic inconsistency. It is thus important that our model of provocation is not committed to one interpretation only.

The origin of provocation matters in at least two ways. First, it may determine the manner in which provocability should be empirically measured or tested. Second, it becomes behaviourally relevant in extensions of our model. Why so? The present model can leave the question open essentially because each individual gets to choose only once a damage-punishment pair, namely after the new policy is implemented, and so only his *then*-preference matters for behaviour, regardless of how it came about. However, in an extended model of repeated interaction between policy maker and population, the origin of provocation affects individual strategies, hence optimal policies to prevent terrorism.

## 2.5 The fallacy of neglecting provocation

Provocation effects of policies are easy to overlook in practice, for systematic reasons given in a moment. Probably they *are* being overlooked or underestimated in contemporary anti-terrorism politics, but this is an empirical claim

[11] Dynamically inconsistent (forward-looking) agents may choose to commit themselves, even if commitment comes with a cost. E.g. Strotz (1955-56) and Hammond (1976).

that this theoretical paper cannot defend. Rather, let me briefly discuss (i) consequences of and (ii) explanations for provocation-neglect.

*Consequence of provocation-neglect: toughness exaggeration.* (The rigorous treatment comes in Section 3.4.) A provocation-neglecting policy maker minimises the wrong objective function since he wrongly predicts people's response to the policy. He assumes people keep their old (status quo) preferences under the new policy: peacefully minded persons stay peacefully minded, brutally minded persons stay equally brutally minded. This leads the provocation-neglecting policy maker to minimise $\mathbf{DE}(t) + Cost(t)$ rather than $\mathbf{DE}(t) + \mathbf{PE}(t) + Cost(t)$, where, as in Section 2.3, $\mathbf{DE}(t)$, $\mathbf{PE}(t)$ and $Cost(t)$ denote the aggregate deterrence effect, the aggregate provocation effect, and the cost of policy $t \in T$, respectively. Since $\mathbf{PE}(t)$ is an increasing function of toughness

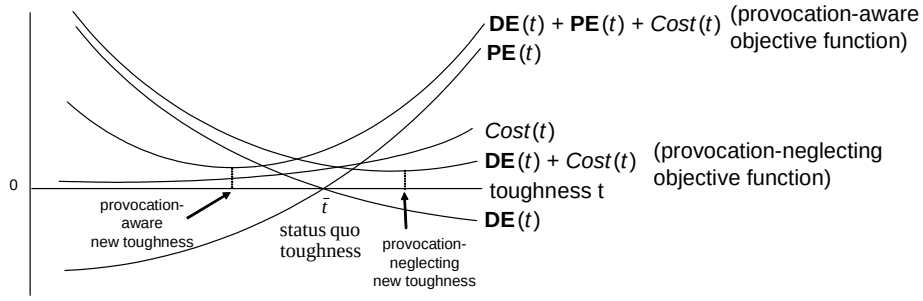


Figure 8: An example in which provocation-neglect leads to a toughness raise, but provocation-awareness to a toughness reduction

(see Theorem 1 below), the provocation-neglecting policy maker is tougher than would be optimal, as Figure 8 illustrates in the case of a one-dimensional policy choice problem. Such toughness exaggeration due to provocation-neglect may be called the 'fallacy of neglecting provocation'.

*Explaining provocation-neglect.* Is provocation-neglect an elementary mistake that cannot be expected to occur among policy makers? On the contrary, provocation-neglect may be a tempting error, almost like a trap wide open in front the policy maker who can avoid it only by particular serenity. The reason is that people's provocability may be little visible *before* the new policy (e.g. before a war): it is not (yet) revealed in behaviour, perhaps not even in speech, especially if provocation comes from dynamic inconsistency (see Section 2.4). In a relatively mild status quo environment, individuals who *would* develop brutal preferences *if* the policy were toughened may display perfectly sane and harmless behaviour, and even declare the peacefulness of their intensions. It certainly

takes special serenity and psychological and cultural sensitivity to foresee if and how a policy would provoke those subgroups who so far behave peacefully and whose (dynamically inconsistent) members are perhaps even themselves unaware of being provocable.

Do there exist indirect ways to nevertheless 'observe' or 'test' beforehand if and how people would be provoked by new policies? In now briefly address this question.

First, consider speech-revelation: can one trust someone's speech about what he *would* desire or do under such and such new policy? Speech-revealed preferences, which many economists legitimately treat with caution, may be particularly unreliable here, for two reasons:

- A person might not sincerely reveal the brutality of his future preferences or actions, by fear of the consequences of making his criminal side known.
- We are dealing here with revealing not the present preference order (over damage-punishment pairs), but future preferences held under potential new policies. Revealing these is *in principle* possible, and might be realistic under the reciprocity interpretation because then the future desires already exist presently in the form of conditional preferences. But the plausibility of revealing future preferences decreases if these preferences arise by dynamic inconsistency (or by rationally acquired tastes): the person might then not be aware that a new environment would turn him into a terrorist, also given that this drastic event would presumably be unprecedented in his life.

Second, while speech-revelation is thus limited as a tool to 'measure' provocability, certain past observations may serve as proxis to estimate provocability. Similar policies might in the past have been used in similar contexts and on similar populations (though many culture- and context-specific factors would have to be controlled for). Also, the population might in the past have displayed certain (more or less violent) behavioural patterns in response to environments (such as more or less rough social environments) that, though not identical to the policy-induced environments, resemble them. In the best case scenario, past data allow one to statistically estimate, for each policy $t \in T$, the population's resulting distribution of preferences $\succeq_t$.

# 3 Formal analysis

The above informal analysis draws on some claims, in particular about the signs of two competing effects of toughness on terrorism: the provocation effect

is typically non-negative, the deterrence effect typically non-positive. But what means 'typically'? I take this question up now by giving sufficient conditions for these claims to hold. A social welfare analysis of toughness will also confirm the earlier claim that provocation-neglect leads to toughness exaggeration.

To draw a comparison first, Slutsky's *fundamental equation of demand theory* decomposes the effect of a price increase on demand into two conceptually distinct effects, the income effect and a substitution effect. We pursue a similar goal in decomposing the effect of toughness on terrorism (damage) into two conceptually distinct effects, deterrence and provocation. In spite of obvious differences, our approach shares some key aspects with Slutsky's:

(i) The primary level of description is the individual: the effect of a price/toughness raise on overall demand/terrorism is obtained by aggregating the effect on individual demand/terrorism. Accordingly, much of this section focusses on a single individual, with the understanding that one could later aggregate.

(ii) The effect of a price/toughness change can be analysed either by comparing the status quo price/toughness with a fixed new price/toughness, as done in textbook discussions of Slutsky's decomposition and in Section 2, or by considering a *marginal* price/toughness change, i.e. by differentiating demand/terrorism with respect to price/toughness, as done in Slutsky's *equation* and in this section.

(iii) Each subeffect has a *typical* sign, yet there are exceptions, such as income effects getting positive for inferior goods and provocation effects getting negative for inversely provocable individuals.

(iv) The two subeffects are constructed by means of a thought experiment involving hypothetical behaviour: the substitution (resp. deterrence) effect represents how a price (resp. toughness) change *would* affect demand (resp. terrorism) if, hypothetically, the individual's achieved utility level (resp. his preference order $\succeq$) did not change.

## 3.1 Framework, terminology, notation

Throughout we consider the following sequence of events. First the policy maker chooses a policy $t$ from a given set $T$ of possible policies, among which the status quo policy $\bar{t}$. Each policy $t \in T$ leads to a set $\mathbf{F}_t \subseteq \mathbb{R}^2_+$ of feasible damage-punishment pairs $(x, y)$. Under the environment of policy $t \in T$, each individual $i$ in the (finite non-empty) population $N$ holds some (complete and transitive) preference order $\succeq_{t,i}$ on the damage-punishment quadrant $\mathbb{R}^2_+$ that guides his behaviour, i.e. his choice of a damage-punishment pair $(x, y)$ from the feasible set $\mathbf{F}_t$. I call the policy maker *provocation-aware* (resp. *-neglecting*)

if he believes that any policy $t \in T$ gives any individual $i$ the true preferences $\succeq_{i,t}$ (resp. the status quo preference $\succeq_{i,\bar{t}}$).

In spite of what the figures in Section 2 might suggest, feasible sets $\mathbf{F}_t$ need neither be linear, nor be 'thin', nor render all damage levels $x \in \mathbb{R}_+$ feasible. In general, each feasible set has two by-products: the *feasible damage set* and the *punishment function*, defined and denoted as follows. Policy $t$'s *feasible damage set* $\mathbf{X}_t \subseteq \mathbb{R}_+$ is defined as $\mathbf{X}_t := \{x : (x, y) \in \mathbf{F}_t \text{ for some } y \in \mathbb{R}_+\}$ (the projection of $\mathbf{F}_t$ on the $x$-coordinate), representing what damage is physically possible under policy $t$. If $\mathbf{X}_t = \mathbb{R}_+$, *any* damage can be created. If $\mathbf{X}_t \subsetneq \mathbb{R}_+$, the policy physically limits the kind or extent of damage people can create, for instance by weapons embargoes or police presence or airport controls, which make major terrorist attacks simply impossible to create, thus leaving only the more 'modest' targets for terrorists. Typically, the feasible damage set $\mathbf{X}_t$ contains at least $x = 0$ ('no damage' is feasible) and forms an interval, which could be unbounded ($\mathbf{X}_t = \mathbb{R}_+$) or bounded ($\mathbf{X}_t = [0, x_t^*]$ or $\mathbf{X}_t = [0, x_t^*)$ where $x_t^*$ is a feasibility bound established by policy $t$). Policy $t$'s *punishment function* $f_t : \mathbf{X}_t \to \mathbb{R}_+$ maps every feasible damage level $x \in \mathbf{X}_t$ to the minimally received punishment, i.e. $f_t(x) := \inf\{y : (x, y) \in \mathbf{F}_t\}$. The graph of the function $f_t$ represents the southern border of the feasible set $\mathbf{F}_t$. An important example are feasible sets of the ('thin') form

$$\mathbf{F}_t = \{(x, f_t(x)) : x \in \mathbf{X}_t\},\ t \in T, \tag{1}$$

consisting of pairs of a feasible damage $x \in \mathbf{X}_t$ and a *unique* punishment $f_t(x)$; here each feasible set $\mathbf{F}_t$ is 'thin' in that it coincides with the graph of $f_t$.

Typically, a policy $t$'s punishment function $f_t$ is increasing: higher punishment for higher damage. Intuitively, the tougher the policy $t$, the higher the punishments $f_t(x)$, $x \in \mathbf{X}_t$, and also the fewer the feasible damage levels, i.e. the smaller $\mathbf{X}_t$. However, it is perfectly possible for two policies in $T$ that one gives higher punishment yet renders more damage levels feasible, or that one gives more punishment for some damage levels but less for others; then these two policies cannot easily be ranked in terms of their toughness or deterrence.

In practice, individuals often have many ways to produce a given damage $x \geq 0$ (e.g. many ways to kill someone), and punishment might depend on the chosen way. Feasible sets do then not take the 'thin' form (1) but the general form

$$\mathbf{F}_t = \{(x, y) \in \mathbb{R}_+^2 : y \in Y_t(x)\},\ t \in T,$$

where, for each policy $t \in T$ and each damage level $x \geq 0$, $Y_t(x)$ is a set $Y_t(x) \subseteq \mathbb{R}_+$ of punishment levels that can occur in combination with damage level $x$. In fact, feasible sets $\mathbf{F}_t$, $t \in T$, can always be written in the latter form;

in the 'thin' case (1), each set $Y_t(x)$ is singleton (if $x$ is feasible) or empty (if $x$ is infeasible).

## 3.2 Marginal provocation effect and deterrence effect

In the rest of Section 3 (but not in Appendix A), the policy maker chooses a single policy parameter representing the toughness level. More precisely:

**Unidimensional Policy Space UP.** The set of policies $T$ is an interval $T \subseteq \mathbb{R}$ (of *toughness levels*), and the status quo $\bar{t} \in T$ is non-extremal, i.e. not on the boundary of the interval $T$.

We consider an individual whose preferences $\succeq_t$, $t \in T$, are *regular* as defined by three conditions:

- **R1** (*punishment-aversion*) For every policy $t \in T$, $\succeq_t$ is punishment-averse, i.e. $(x, y) \succ_t (x, y')$ whenever $y < y'$.
- **R2** (*continuity*) For every policy $t \in T$, $\succeq_t$ is continuous, hence is (by Debreu's Theorem) representable by a continuous *utility* function $u_t : \mathbb{R}^2_+ \to \mathbb{R}$.
- **R3** (*unique optimum*) For all policies $t_1, t_2 \in T$, there exists a unique damage-punishment pair, denoted $(\mathrm{x}(t_1, t_2), \mathrm{y}(t_1, t_2))$, that maximises $\succeq_{t_1}$ within $\mathbf{F}_{t_2}$ (i.e. that is an optimal response to policy $t_2$ under the preferences of policy $t_1$), and moreover the damage level $\mathrm{x}(t_1, t_2)$ is a differentiable function of $(t_1, t_2) \in T \times T$.[12]

The optimisation problem in R3 is hypothetical in that under policy $t_2$ the individual really maximises $\succeq_{t_2}$, not $\succeq_{t_1}$; but if we set $t_1 = t_2 = t$, we obtain precisely the individual's real optimisation problem under policy $t$. Hence, R3 in particular implies that

- to each policy $t \in T$ the individual has a unique optimal response, to be denoted $(\mathrm{x}(t), \mathrm{y}(y))$ $(= (\mathrm{x}(t, t), \mathrm{y}(t, t)))$.

So we can define the marginal effect of raising toughness from the status quo $\bar{t}$:

- The *(marginal) effect (of toughness)* is defined as $E := \mathrm{x}'(\bar{t})$, the derivative (at the status quo) of damage with respect to toughness.[13]

How can we meaningfully decompose $E$ into two subeffects? As illustrated

[12] Throughout, the derivative of a function at a point on the boundary of the function's domain is interpreted as usual, i.e. as a one-sided derivative.

[13] R3 ensures that $\mathrm{x}(t)$ (and $\mathrm{x}_{\mathrm{deter}}(t)$ and $\mathrm{x}_{\mathrm{prov}}(t)$ defined below) are indeed differentiable functions. See the proof of Theorem 1.

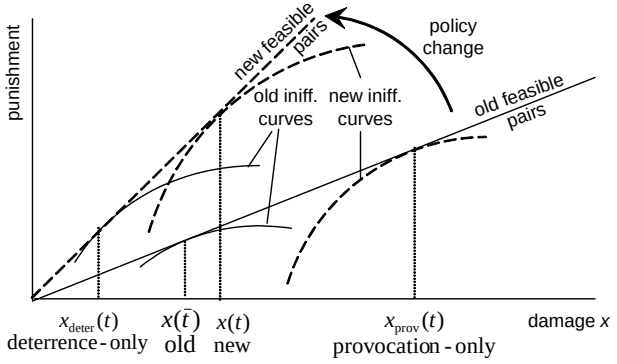


Figure 9: The old, the new, and the two hypothetical damage levels

in Figure 9, the key is to first introduce two hypothetical behaviours, one that neglects provocation and one that neglects deterrence:

- The *pure-deterrence* or *provocation-neglecting* response to policy $t \in T$, denoted $(\mathrm{x}_{\mathrm{deter}}(t), \mathrm{y}_{\mathrm{deter}}(t))$, is defined as $(\mathrm{x}(\bar{t}, t), \mathrm{y}(\bar{t}, t))$, the choice that maximises the old preference $\succeq_{\bar{t}}$ within the new feasible set $\mathbf{F}_t$. It captures deterrence without provocation, as it represents how the individual would react to the new policy if (hypothetically) his preferences were to remain unchanged. It represents how a provocation-neglecting policy maker predicts the individual's response to policy $t$.
- The *pure-provocation* or *deterrence-neglecting* response to policy $t \in T$, denoted $(\mathrm{x}_{\mathrm{prov}}(t), \mathrm{y}_{\mathrm{prov}}(t))$, is defined as $(\mathrm{x}(t, \bar{t}), \mathrm{y}(t, \bar{t}))$, the choice that maximises the new preference $\succeq_t$ within the old feasible set $\mathbf{F}_{\bar{t}}$. It captures provocation without deterrence, by representing how the individual would react to the new policy $t$ if (hypothetically) he did not yet face the new constraints (such as new punishment levels). Under another interpretation, it represents the individual's reaction if, although his preferences are already affected (perhaps provoked) by the new environment, he is short-sighted or irrational in that he ignores the new punishment levels he faces.

The pure-deterrence damage $\mathrm{x}_{\mathrm{deter}}(t)$ and pure-provocation damage $\mathrm{x}_{\mathrm{prov}}(t)$ represent two partial views on the person's damage response to policy $t$: $\mathrm{x}_{\mathrm{deter}}(t)$ is optimal under old preferences given new punishment levels, and $\mathrm{x}_{\mathrm{prov}}(t)$ is optimal under new preferences supposing old punishment levels. I can now define deterrence and provocation effects.

- The *(marginal) deterrence effect (of toughness)* is defined as $DE := \mathrm{x}'_{\mathrm{deter}}(\bar{t})$, the derivative (taken at the status quo $t = \bar{t}$) of the pure-deterrence damage $\mathrm{x}_{\mathrm{deter}}(t)$. It captures the marginal damage change as far as it is caused

by changing constraints (punishments), ignoring any meanwhile preferences change.

- The *(marginal) provocation effect (of toughness)* is defined as $PE := \mathrm{x}'_{\mathrm{prov}}(\bar{t})$, the derivative (taken at the status quo $t = \bar{t}$) of the pure-provocation damage $\mathrm{x}_{\mathrm{prov}}(t)$. It captures the marginal damage change as far it is caused by preference change, ignoring the changing constraints (punishments).

## 3.3 Theorem

I now show that the total effect of toughness is decomposable into $E = PE + DE$ with $PE \geq 0$ and $DE \leq 0$. While the additive decomposition $E = PE + DE$ is simple to prove (essentially, by applying the chain rule), the claim on the signs of the subeffects is non-trivial and does not hold universally, but under meaningful conditions. Specifically, each inequality, $PE \geq 0$ and $DE \leq 0$, is based on exactly one condition on preferences. The condition for $PE \geq 0$ excludes that raising toughness appeases the preference; more precisely:

**Condition NIP** (*no inverse provocability*) If the individual is currently indifferent between two damage-punishment pairs, then a toughness raise cannot make him prefer the pair with lower damage. That is, whenever $(x, y) \sim_{\bar{t}} (x', y')$ with $x < x'$, then no policy $t \in T$ with $t > \bar{t}$ leads to $(x, y) \succ_t (x', y')$.

NIP is plausible – it is less demanding than requiring that $\succeq_t$ be *at least as brutal* as $\succeq_{\bar{t}}$ (see Section 2.1) whenever $t > \bar{t}$ – but not universal: surely, there also exist inversely provocable individuals, such as ones who after the traumatic experience of a war lose any desire to exercise violence themselves. For such individuals, the provocation effect $PE$ can become negative.

As the deterrence effect $DE$ is (unlike $PE$) defined by holding preferences fixed, the inequality $DE \leq 0$ has to be based on a condition quite different to NIP: not a condition about how preference changes as the policy changes, but one about internal consistency of status quo preference:

**Condition TC** (*translation-consistency*) Under the status quo preferences, a preference of one damage-punishment pair over another with higher punishment is not reversed by any symmetric punishment increase. That is, whenever $(x, y) \succ_{\bar{t}} (x', y')$ with $y < y'$, then for no $\epsilon > 0$ there is $(x, y + \epsilon) \prec_{\bar{t}} (x', y' + \epsilon)$.

By TC, an extra amount of punishment cannot hurt less if it comes on top of more punishment; for instance, an extra hour of compulsory labour cannot

hurt less if it comes on top of 10 hours than if it comes on top of 5 hours. TC is again plausible but not universal, and its failure can render the deterrence effect $DE$ positive.

Unlike the definition of $PE$, that of $DE$ is based on varying the feasible set $\mathbf{F}_t$, and so the sign of $DE$ cannot possibly be independent of how $\mathbf{F}_t$ reacts to the policy $t \in T$. This is why the inequality $DE \leq 0$ requires an extra condition on feasible sets, one that relates the shape of $\mathbf{F}_t$ to the toughness level $t \in T$. Specifically, I require that the tougher the policy $t \in T$ is, the 'steeper' the feasible set $\mathbf{F}_t$ becomes, i.e. the larger *marginal* punishment becomes:

**Condition MP** (*marginal punishment increases with toughness*) Each feasible set $\mathbf{F}_t$, $t \in T$, contains the no-damage-no-punishment pair $(0, 0)$, it is (topologically) closed and connected, and its marginal punishment function $f'_t : \mathbf{X}_t \to \mathbb{R}$ is defined[14], non-negative, and (at least weakly) increasing in toughness $t$[15].

Essentially, MP requires the southern border of the feasible set $\mathbf{F}_t$ (i.e. the graph of $f_t$) to have a non-negative slope that increases if toughness $t$ increases. In the special case that each feasible set $\mathbf{F}_t$ is 'thin' (i.e. identical to its southern border: $\mathbf{F}_t = \{(x, f_t(x)) : x \in \mathbf{X}_t\}$), MP simply requires feasible sets to everywhere have a non-negative slope that increases with toughness. MP holds for instance if $T = (0, \infty)$ and each $\mathbf{F}_t$ has southern border of

- the linear form $f_t(x) = tx$ (so $f'_t(x) = t$), or more generally,
- the form $f_t(x) = tx^c$ for a fixed $c > 0$ (so $f'_t(x) = ctx^{c-1}$),

because $f'_t$ is then non-negative and increasing in $t$.

**Theorem 1** *Consider the unidimensional policy choice problem UP. Then, for every individual whose preferences are regular (i.e. satisfy R1-R3),*

(a) *the effect of toughness on terrorism is the sum of the provocation and deterrence effects: $E = PE + DE$;*

(b) *the two subeffects are opposed, that is:*
- *$PE \geq 0$ if individual preferences satisfy NIP;*
- *$DE \leq 0$ if individual preferences satisfy TC and policies satisfy MP.*

This theorem (proved in Appendix B) confirms Section 2's analysis of a trade-off between deterrence and provocation, this time from a *marginal* toughness angle, i.e. from a comparative statics angle. Indeed, under Theorem 1's

[14] That is, the punishment function $f_t : \mathbf{X}_t \to \mathbb{R}_+$ is differentiable.

[15] That is, $t < t'$ implies that $f'_t(x) \leq f'_{t'}(x)$ at all damage levels $x \geq 0$, with $f'_t(x)$ (resp. $f'_{t'}(x)$) naturally read as $\infty$ if $x$ is infeasible, i.e. if $x \notin \mathbf{X}_t$ (resp. $x \notin \mathbf{X}_{t'}$).

conditions the two subeffects pull in opposite directions, and whether a marginal toughness rise increases terrorism by the person depends on which of $PE$ and $DE$ dominates. Arguably, this comparison is what policy makers should mainly focus on in practice.

Of course, Theorem 1 implies an analogous decomposition at the aggregate level: $\mathbf{E} = \mathbf{PE} + \mathbf{DE}$, with $\mathbf{E}$, $\mathbf{PE}$, and $\mathbf{PE}$ defined as the sum-total of the individual effects $E$, $PE$ and $DE$ across the population, respectively.

## 3.4 The social utility of toughness

So far I have been largely informal about the policy maker's preferences, occasionally assuming that he minimises sum-total terrorism or sum-total terrorism plus policy costs. More generally, assume now he holds some arbitrary preference order over the set $\mathbb{R}_+^N \times T$ of damages-policy combinations $((x_i)_{i\in N}, t)$, and let this preference be representable by a 'social utility' function $U : \mathbb{R}_+^N \times T \to \mathbb{R}$ such that

- (*terrorism-aversion*) $U$ is an (at least weakly) decreasing function of each individual $i$'s damage level $x_i \in \mathbb{R}_+$.

There are numerous examples. Social utility may be defined by $U((x_i)_{i\in N}, t) = -\sum_{i\in N} x_i$ if the policy maker minimises sum-total terrorism, or by $U((x_i)_{i\in N}, t) = \#\{i \in N : x_i = 0\}$ if he minimises the number of terrorists (individuals with positive damage). A more general specification is $U((x_i)_{i\in N}, t) = -\sum_{i\in N} x_i^\alpha$ (with a fixed parameter $\alpha > 0$), which reduces to the first example if $a = 1$ and to the second one if $\alpha \to 0$. Another natural class of utility functions are the Cobb-Douglas forms $U((x_i)_{i\in N}, t) = -\Pi_{i\in N} x_i^\alpha$ (for some parameter $\alpha > 0$). Each of these specifications can be refined by subtracting a (suitably scaled) cost term $Cost(t)$ that captures financial costs or other negative policy effects such as loss of (civilian or military) lives or loss of life quality through more state supervision; this gives for instance the utility specification $U((x_i)_{i\in N}, t) = -\sum_{i\in N} x_i - Cost(t)$. Note that by *subtracting* a cost term we assume that policy costs are additively separable from the damage disutility. A specification without additive separability is $U((x_i)_{i\in N}, t) = -t^\beta \Pi_{i\in N} x_i^\alpha$ (for fixed parameters $\alpha, \beta > 0$), assuming here that $t$ is a toughness level taken from a policy interval $T \subseteq (0, \infty)$ and that toughness is costly (i.e. $\beta > 0$).

We now proceed to a comparative statics analysis that makes the following assumptions. As in the last two subsections, we consider the unidimensional policy choice problem UP, and assume that individual preferences on the damage-punishment quadrant $\mathbb{R}_+^2$ are regular (i.e. satisfy R1-R3), so that each individual $i$ has to any toughness level $t \in T$ a unique optimal damage response

$\mathrm{x}_i(t)$, which is differentiable in $t$. Further, let the utility function $U$ be differentiable. Then the marginal value of toughness $t$ can be captured by the total derivative $dU/dt$ (evaluated at the status quo $\bar{t}$). By the chain rule,

$$\underbrace{\frac{dU}{dt}}_{\substack{\text{marginal}\\ \text{utility of}\\ \text{toughness}}} = \sum_{i \in N} \underbrace{\frac{\partial U}{\partial x_i}}_{\substack{\leq 0\\ \text{marginal}\\ \text{utility of } i\text{'s}\\ \text{damage}}} \underbrace{\frac{d\mathrm{x}_i}{dt}}_{\substack{\leq 0 \text{ or } \geq 0\\ \text{individual}\\ \text{damage}\\ \text{response}}} + \underbrace{\frac{\partial U}{\partial t}}_{\substack{\leq 0\\ \text{cost}\\ \text{effect}}}. \tag{2}$$

So the marginal utility of toughness is composed of a (direct) cost effect $\frac{\partial U}{\partial t}$ and (indirect) effects $\frac{\partial U}{\partial x_i}\frac{d\mathrm{x}_i}{dt}$ $(i \in N)$ through people's responses. While the cost effect is typically negative because toughness is expensive, the indirect effects can go in either direction because the sign of the damage response $\frac{d\mathrm{x}_i}{dt}$ may differ across individuals $i$. Using Theorem 1, we can decompose the damage response into the sum $\frac{d\mathrm{x}_i}{dt} = DE_i + PE_i$ of the deterrence effect $DE_i$ and the provocation effect $PE_i$ on individual $i$'s damage, where typically $DE_i \leq 0$ and $PE_i \geq 0$. So, (2) becomes

$$\underbrace{\frac{dU}{dt}}_{\substack{\text{marginal}\\ \text{utility of}\\ \text{toughness}}} = \underbrace{\sum_{i \in N} \frac{\partial U}{\partial x_i} DE_i}_{\substack{\geq 0\\ \text{aggregate deterrence}\\ \text{effect on utility}}} + \underbrace{\sum_{i \in N} \frac{\partial U}{\partial x_i} PE_i}_{\substack{\leq 0\\ \text{aggregate provocation}\\ \text{effect on utility}}} + \underbrace{\frac{\partial U}{\partial t}}_{\substack{\leq 0\\ \text{cost}\\ \text{effect}}}. \tag{3}$$

Three competing forces thus act on the marginal utility of toughness: $\frac{dU}{dt}$ is increased by an aggregate deterrence term, but decreased both by an aggregate provocation term and the cost effect. Whether a toughness increase is desirable depends on whether the deterrence term overweighs the two other terms.

Suppose further the policy maker does not care about *who* creates damage, in the sense that $U((x_i)_{i \in N}, t) = U((x'_i)_{i \in N}, t)$ whenever the damage profiles $(x_i)_{i \in N}$ and $(x'_i)_{i \in N}$ display identical total damage $\sum_{i \in N} x_i = \sum_{i \in N} x'_i$. As one easily shows, the partial derivative $\frac{\partial U}{\partial x_i}$ is then the same for each individual $i$; so it can be bracketed out in (3), and we obtain

$$\underbrace{\frac{dU}{dt}}_{\substack{\text{marginal}\\ \text{utility of}\\ \text{toughness}}} = \underbrace{\frac{\partial U}{\partial x_i}}_{\substack{\leq 0\\ \text{marginal}\\ \text{utility of}\\ \text{terrorism}}} \Big( \underbrace{\mathbf{DE}}_{\substack{\leq 0\\ \text{aggregate}\\ \text{deterrence}\\ \text{effect}}} + \underbrace{\mathbf{PE}}_{\substack{\geq 0\\ \text{aggregate}\\ \text{provocation}\\ \text{effect}}} \Big) + \underbrace{\frac{\partial U}{\partial t}}_{\substack{\leq 0\\ \text{cost}\\ \text{effect}}}, \tag{4}$$

where **DE** and **PE** are, as usual, the aggregate deterrence effect $\sum_{i \in N} DE_i$ resp. provocation effect $\sum_{i \in N} PE_i$. Whether raising toughness is beneficial

depends on the sign of $\frac{dU}{dt}$. It is beneficial if $\frac{dU}{dt} > 0$, which (assuming that $\frac{\partial U}{\partial x_i}$ is *strictly* negative) happens exactly when

$$\underbrace{\mathbf{DE}}_{\leq 0} + \underbrace{\mathbf{PE}}_{\geq 0} < \underbrace{-\frac{\partial U/\partial t}{\partial U/\partial x_i}}_{\leq 0}, \tag{5}$$

i.e. when the two competing effects, **DE** and **PE**, are overall 'sufficiently negative'. By contrast, a provocation-neglecting policy maker (who believes that preferences are policy-invariant, hence that $\mathbf{PE} = 0$) raises toughness already if

$$\mathbf{DE} < -\frac{\partial U/\partial t}{\partial U/\partial x_i}, \tag{6}$$

hence more easily because $\mathbf{PE} \geq 0$. The criteria (5) and (6) illustrate the behavioural difference between accounting for and neglecting provocation: in his decision over whether to raise toughness, the provocation-aware policy maker is guided by the more restrictive criterion (5), hence raises toughness less easily and reduces toughness more easily. The equilibrium toughness level (at which the policy maker neither raise nor reduces toughness) is lower for the provocation-aware policy maker, because his first-order condition

$$\mathbf{DE} + \mathbf{PE} = -\frac{\partial U/\partial t}{\partial U/\partial x_i} \tag{7}$$

is typically satisfied at a lower status quo toughness than the provocation-neglecting policy maker's first-order condition

$$\mathbf{DE} = -\frac{\partial U/\partial t}{\partial U/\partial x_i}. \tag{8}$$

The different equilibrium conditions (7) and (8) again illustrate the different toughness dispositions underlying provocation-aware and -neglecting policy making.

# 4 Concluding remarks

First, let me summarise some issues that have been developed. A policy maker may try to reduce terrorism either by cause-related measures, which aim to appease people's preferences, or by symptom-related measures, which change the constraints (feasible set) under which terrorists operate and which, importantly, may have side effects on some people's preferences ('provocation'). While both approaches are costly, a fundamental difference lies in the benefit

side: cause-related measures reduce terrorism (the only question being: by how much?), whereas symptom-related measures may or may not reduce terrorism, depending on whether the deterrence effect overweighs the provocation effect. This does not imply that symptom-related measures are generally inferior, but that they bear a higher downside-risk: the worst outcome of a cause-related measure is to incur the policy cost without terrorism reduction, but the worst outcome of a symptom-related measure (e.g. a war) is to incur the policy cost with a terrorism increase. Most of the paper has focussed on analysing symptom-related policies, and specifically the trade-off between deterrence and provocation. Theorem 1 provides general sufficient conditions under which the marginal deterrence effect is non-positive (i.e. terrorism-reducing) and the marginal provocation effect is non-negative (i.e. terrorism-increasing). I have argued that it is easy in practice to overlook provocation effects (Section 2.5). Provocation-neglect leads to toughness exaggeration (the *fallacy of neglecting provocation*), as argued informally in Section 2.5 and shown formally in Section 3.4 by comparing the (first-order) conditions under which a provocation-aware and a provocation-neglecting policy maker chooses how tough to be. As an analytic example, Theorem 2 in Appendix A characterises optimal policies for a stylised objective function (minimising the number of terrorists), again confirming that there is a trade-off between deterring and provoking, and that provocation-neglect leads to toughness exaggeration.

Our analysis poses several empirical and theoretical challenges. On the empirical side, it would be of high practical interest to know the extent to which concrete cause-related policies (investments into social stability, into education etc.) appease preferences,[16] and the extent to which concrete symptom-related policies (weapons embargoes, military presence, criminal legislation etc.) brutalise preferences, i.e. provoke. Among different ways to deter (i.e. to render the feasible set $\mathbf{F} \subseteq \mathbb{R}^2_+$ 'steeper'), which ones provoke least? Are there policies that deter without provoking? A concrete hypothesis to investigate is whether defensive deterrence provokes less than aggressive deterrence. As mentioned in Section 2.2, I count a deterrence measure as defensive if it makes it harder to create terrorism (e.g. weapons embargoes) and as aggressive if it increases punishment (e.g. tough criminal legislation).[17] To provide answers to such

[16] This question relates to existing empirical research apart from the difference between appeasing preferences and reducing terrorism.

[17] Both kinds of deterrence can be defined formally and do indeed constitute deterrence, i.e. render the feasible set $\mathbf{F} \subseteq \mathbb{R}^2_+$ 'steeper'. Let $A$ be a policy-independent set of possible (modes of) behaviour, and let $\mathrm{x}(a,t) \in \mathbb{R}_+$ resp. $\mathrm{y}(a,t) \in \mathbb{R}_+$ be the damage resp. punishment resulting from behaviour $a \in A$ at policy $t \in T$. So policy $t \in T$ leads to the feasible set $\mathbf{F}_t = \{(\mathrm{x}(a,t), \mathrm{y}(a,t)) : a \in A\} \subseteq \mathbb{R}^2_+$. Say that one policy does more defensive (resp. aggressive) deterrence than another if it implies lower damage $\mathrm{x}(a,t)$ (resp. higher punishment

questions, it would help to understand the psychological phenomenon of provocation: when is it a form of reciprocity, when one of taste acquisition, and when one of dynamic inconsistency (see Section 2.4)?

On a theoretical dimension, there is plenty of room for adapting our model to concrete applications, or for 'merging' it with models studied in political economy (that is, incorporating provocation into these models), or for refining the strategic interaction between individuals and the policy maker. For instance, one might introduce uncertainty of the policy maker about the types of population members, i.e., about how brutal and how provocable their preferences are. Or, one might model a repeated interaction between policy maker and population, with the question arising as to whether toughness can provoke only in the short run or can have lasting provocation effects on the population.

# 5 References


**Becker, G.** 1996. *Accounting For Tastes*. Harward University Press, Cambridge Mass., London.

**Bénabou, R., Pycia, M.** 2002. "Dynamic inconsistency and self-control: a planner-doer interpretation." *Economic Letters*, 77: 419-424.

**Bolton, G., Ockenfels, A.** 2000. "ERC: a theory of equity, reciprocity and competition." *American Economic Review*, 90: 166-193.

**Bowles, S.** 1998. "Endogenous preferences: the cultural consequences of markets and other economic institutions." *Journal of Economic Literature*, 36(1): 75-111.

**Brams, S., Kilgour, M.** 1985. "The Path to Stable Deterrence." In: U. Luterbacher/M.D. Ward (eds.), *Dynamic Models of International Conflict*, Boulder/C.O., 11-25.

**Brams, S., Kilgour, M.** 1987. "Is Nuclear Deterrence Rational, and Will Star Wars Help?" *Analyse und Kritik*, 9: 62-74.

**Brams, S., Kilgour, M.** 1988. *Game Theory and National Security*, New York: Basil Blackwell.

**Bueno de Mesquita, E.** 2005. "The Quality of Terror." *American Journal of Political Science*, 49(3): 515-530.

**Cioffi-Revilla, C.** 1998. "On the likely magnitude, extent, and duration of an Iraq-UN war." *Journal of Conflict Resolution*, 35: 387-411.


$\mathrm{y}(a,t)$) for each individual behaviour $a \in A$. Installing a (functioning) missile defense system, for instance, constitutes defensive deterrence: the action $a$ of sending a missile would not anymore create damage.

**Cioffi-Revilla, C.** 1985. "Political reliability theory and war in the International system." *American Journal of Political Science*, 29: 47-68.

**Dekel, E., Ely, J., Yilankaya, O.** 2007. "Evolution of Preferences." *The Review of Economic Studies*, 74: 685-704.

**Dietrich, F.** 2012. "Modelling change in individual characteristics: an axiomatic framework." *Games and Economic Behavior* 76, 471-94

**Dufwenberg, M., Kirchsteiger, G.** 2004. "A theory of sequential reciprocity." *Games and Economic Behavior*, 47(2): 268-298.

**Dumas, L.** 2002. "Is development an effective way to fight terrorism?" *Philosophy & Public Policy Quarterly*, 22(4): 7-12.

**Enders, W., Sandler, T.** 2006. *The political Economy of Terrorism.* Cambridge University Press.

**Eubank, W., Weinberg, L.** 1994. "Does democracy encourage terrorism?" *Terrorism and Political Violence*, 6(4): 155-64.

**Falk, A., Fischbacher, U.** 2006. "A Theory of Reciprocity." *Games and Economic Behavior*, 54 (2): 293-315.

**Fehr, E., Gächter, S.** 1998. "Reciprocity and economics: the economic implications of homo reciprocans." *European Economic Review*, 42: 845-859.

**Frey, B.** 2004. *Dealing with Terrorism: Stick or Carrot?* Edward Elgar Publishing Ltd., Cheltenham, UK and Northampton, Mass.

**Frey, B. S., Luechinger, S.** 2003. "How to fight terrorism: Alternatives to deterrence." *Defence and Peace Economics*, 14(4): 237-249.

**Goodin, R. E. 2006.** *What's Wrong with Terrorism?* Cambridge: Polity Press

**Hammond, P.** 1976. "Changing Tastes and Coherent Dynamic Choice." *Review of Economic Studies*, 43: 159–173.

**Hansson, S. O.** 1995. "Changes in preference." *Theory and Decision*, 38: 1-28.

**Jehle, G., Reny, P.** 2001. *Advanced Microeconomic Theory.* 2nd Edition. International Edition.

**Krueger, A. B., Malecková, J.** 2003. "Education, poverty and terrorism: Is there a causal connection?" *Journal of Economic Perspectives*, 17(4): 119-144.

**Lichbach, M. I.** 1987. "Deterrence or escalation? The puzzle of aggregate studies of repression and dissent." *Journal of Conflict Ressolution*, 31(2): 266-297.

**O'Donoghue, E., Rabin, M.** 1999. "Doing it now or doing it later." *American Economic Review*, 89: 103-124.

**Polak, R.** 1976. "Interdependent preferences." *American Economic Review*

66(3): 309-20.

**Rabin, M.** 1993. "Incorporating fairness into game theory and economics." *American Economic Review*, 83: 1281-1302.

**Rabin, M.** 1998. "Psychology and economics." *Journal of Economic Literature*, 36(1): 11-46.

**Rosendorff, P., Sandler, T.** 2004. "Too Much of a Good Thing? The Proactive Response Dilemma." *Journal of Conflict Resolution*, 48(4): 657-671.

**Sethi, R., Somanathan, E.** 2001. "Preference evolution and reciprocity." *Journal of Economic Theory*, 97: 273-297.

**Sethi, R., Somanathan, E.** 2003. "Understanding Reciprocity." *Journal of Economic Behavior and Organization*, 50(1): 1-27.

**Silke, A.** 1994. *Research on Terrorism: Trends, Achievements and Failures*. Frank Cass Publishers, New York.

**Simon, C., Blume, L.** 1994. *Mathematics for Economists*. Norton & Co. Inc., New York.

**Strotz, R.** 1955-56. "Myopia and inconsistency in dynamic utility maximization." *Review of Economic Studies*, 23(3): 165-180.

**Zagare, F., Kilgour, M.** 2000. *Perfect Deterrence*. Cambridge University Press.

# A An example: minimising the number of terrorists

Unlike in much of this paper, policies may now contain both cause- and symptom-related measures; they might be represented as vectors $t = (t_1, ..., t_k)$ of positions on $k$ cause- or symptom-related dimensions. While I have so far ensured high generality of preferences by placing no specific restrictions (except from plausible ones such as punishment-aversion of individuals and terrorism-aversion of the policy maker), let us now turn to concrete preferences. I take the policy maker to follow a paradigmatic and simple objective – minimising the number of terrorists – and the individuals to hold preferences from a plausible but special class. This will allow us to *analytically* determine optimal policies. The upshot will be that optimal policies again have to strike the right compromise between provocation and deterrence, albeit in a particular sense.

Specifically, individual preferences fall into the following class.

**Preference Model PM**. Each individual in the population $N$ is of one of the

following types.

*Either* he has a *peaceful* type, meaning that under each policy $t \in T$ his preference order $\succeq_t$ (on $\mathbb{R}^2_+$) is peaceful as defined Section 2.1. (The set of peaceful types can be defined as the set of families $(\succeq_t)_{t\in T}$ of peaceful preference orders.)

*Or* his type belongs to the set $\mathbb{R}$ of *possibly brutal* types. Each possibly brutal type $\theta$ in $\mathbb{R}$ ($\theta$ represents the *status quo damage-inclination*) holds under every policy $t \in T$ a preference order $\succeq_t = \succeq_{t,\theta}$ that is punishment-averse (see Section 2.1), strictly convex[18], and representable by a *utility* function $u_t : \mathbb{R}^2_+ \to \mathbb{R}$ of the (separable) form

$$u_t(x, y) = v_t(x) - d(y) \text{ (for all } x, y \geq 0)$$

with the following interpretation and properties:

- $d(y)$ represents *disutility from punishment*, where the function $d : \mathbb{R}_+ \to \mathbb{R}$ is policy-independent (a plausible restriction) and differentiable.
- $v_t(x)$ represents *utility of damage* and takes the form $v_t(x) = v(x - \theta - P(t))$, where:
  - $v : \mathbb{R} \to \mathbb{R}$ is a differentiable and strictly concave function that peaks at 0; hence $v_t(x)$ peaks at $x = \theta + P(t)$, and so the type has a brutal preference with preferred damage level $\theta + P(t)$ under policies $t$ with $\theta + P(t) > 0$, and peaceful preferences under policies $t$ with $\theta + P(t) \leq 0$.
  - $P(t) \in \mathbb{R}$ is interpreted as the amount by which policy $t$ provokes[19], i.e. increases the preferred damage level from the status quo $\bar{t}$, and accordingly I assume without loss of generality that $P(\bar{t}) = 0$;[20]
  - $\theta$ is interpreted as the *status quo damage-inclination*, as it is the status quo preferred damage level (by $P(\bar{t}) = 0$) provided $\theta \geq 0$.
- $v$, $d$ and $P(t)$ are the same across types $\theta \in \mathbb{R}$ (this is the main restriction, essential for analytic tractability).
- There is no policy $t \in T$ at which all types $\theta \in \mathbb{R}$ most prefer the no-damage-no-punishment pair $(x, y) = (0, 0)$ from $\mathbf{F}_t$ (this excludes trivial

[18] That is, for every $(x, y) \in \mathbb{R}^2_+$ the upper contour set $\{(x', y') \in \mathbb{R}^2_+ : (x', y') \succeq_t (x, y)\}$ is strictly convex. It follows that $u_t$ is quasi-concave (Th. A1.14 in Geoffrey Jehle and Philip Reny 2001; to be precise, this theorem uses a slightly stronger notion of strictly convex preferences, which is implied by ours given that we also assume $\succeq_t$ to be continuous and punishment-averse).

[19] The more general term "affects preference" is perhaps better here than "provokes", as $t$ could contain cause-related measures (aimed at preference appeasement, i.e. at $P(t) < 0$).

[20] One may always achieve $P(\bar{t}) = 0$ by subtracting $P(\bar{t})$ from each $P(t)$, $t \in T$, while adding $P(\bar{t})$ to each individual's type $\theta \in \mathbb{R}$; this normalisation leaves individual preferences unchanged.

solutions to the problem of minimising the number of terrorists).

PM is a flexible model: the precise forms of $v, d, P, T$ can be chosen to match the intended application. As as simple example, let the policy space be unidimensional, say the interval $T = \mathbb{R}_+$ of toughness levels, let disutility of punishment be linear, i.e. given by $d(y) = by$ for a fixed parameter $b > 0$, let provocation be also linear, i.e. given by $P(t) = (t - \bar{t})p$ for a fixed parameter $p > 0$ (recall that $P(t)$ represents the change of preferred damage level if toughness changes from the status quo level $\bar{t}$ to $t$), and let the function $v$ be given by $v(x) := -|x|^a$ for a fixed parameter $a > 2$. In summary, then, under toughness $t \in \mathbb{R}_+$ type $\theta \in \mathbb{R}$ has the (quasi-linear) utility function

$$u_t(x, y) = -|x - \theta - (t - \bar{t})p|^a - by \text{ for all } (x, y) \in \mathbb{R}^2_+,$$

with preferred damage level given by $\theta + (t - \bar{t})p$ (or by 0 if this number is $< 0$).

While the policy space $T$ is arbitrary (perhaps multi-dimensional with cause- and symptom-related dimensions), a convexity property is required:

**Convex Punishment CP.** For each policy $t \in T$, the feasible set $\mathbf{F}_t$ contains the no-damage-no-punishment pair $(0, 0)$, it is (topologically) closed and connected, and its punishment function $f_t : \mathbf{X}_t \to \mathbb{R}_+$ is weakly convex.

Recall that $f_t(x)$ represents the minimal punishment for damage $x \in \mathbf{X}_t$; the graph of $f_t$ is the southern border of the feasible set $\mathbf{F}_t$. By CP, this southern border has a weakly convex shape, for instance a linear shape given by $f_t(x) = \alpha_t x$ for some policy-dependent slope $\alpha_t \geq 0$. In general, as $\mathbf{F}_t$ is connected and contains $(0, 0)$, the feasible damage set $\mathbf{X}_t \subseteq \mathbb{R}_+$ is an interval containing 0, hence is either $\mathbb{R}_+$ (unlimited feasibility) or of the form $[0, x_t^*]$ or $[0, x_t^*)$ (with a finite feasibility bound $x_t^*$).

**Theorem 2** *Assuming the preference model PM and convex punishment CP,*

- ***(a)*** *the expression $D(t) := v'^{-1}(d'(0)f_t'(0))$ is for each policy $t \in T$ well-defined, i.e. the (right hand) derivative $f_t'(0)$ exists and $d'(0)f_t'(0)$ has a unique inverse image under the derivative function $v' : \mathbb{R} \to \mathbb{R}$;*
- ***(b)*** (provocation-aware policies) *each policy $t \in T$ that minimises $P(t) + D(t)$ minimises the number of terrorists if each individual's response to each policy $t \in T$ maximises his preference within the feasible set $\mathbf{F}_t$;*
- ***(c)*** (provocation-neglecting policies) *each policy $t \in T$ that minimises $D(t)$ minimises the number of terrorists if each individual's response to each policy $t \in T$ maximises his* status quo *preference within the feasible set $\mathbf{F}_t$;*

***(d)*** *the individual responses assumed in (b) and (c) exist and are unique.*

Taking the example given after the definition of PM and assuming each toughness level $t \in T = \mathbb{R}_+$ leads to the (maximal) feasible damage set $\mathbf{X}_t = \mathbb{R}_+$ and to a linear punishment function given by $f_t(x) = tx$ (which becomes 'steeper' if toughness $t$ increases), one finds that $D(t) = -(tb/a)^{1/(a-1)}$, and by minimising $P(t) + D(t) = (t - \bar{t})p - (tb/a)^{1/(a-1)}$ one finds the

$$\textit{optimal toughness level: } t = \frac{(b/a)^{1/(a-2)}}{(p(a-1))^{(a-1)/(a-2)}}$$

(as derived at the end of Appendix B). So optimal toughness increases if marginal provocation $p = P'(t)$ falls (i.e. if toughness provokes less), and also if marginal punishment aversion $b = d'(y)$ increases (i.e. if punishment hurts more, hence deters more). These comparative statics confirm our intuition.

Theorem 2 once again confirms the trade-off between provocation and deterrence: $P(t)$ is a measure for how much the policy $t$ provokes, and $D(t)$ is an (inverse) measure for how much $t$ deters. $D(t)$ measures deterrence in that it reflects the punishment function $f_t$ but not any policy-induced preference change. $D(t)$ measures deterrence *inversely*: the more deterring $t$, the 'steeper' the feasible set $\mathbf{F}_t$, hence the higher the derivative $f_t'(0)$, so the higher the product $d'(0)f_t'(0)$, and therefore, the lower $D(t) = v'^{-1}(d'(0)f_t'(0))$ because we have applied a strictly decreasing function $v'^{-1}$.[21]

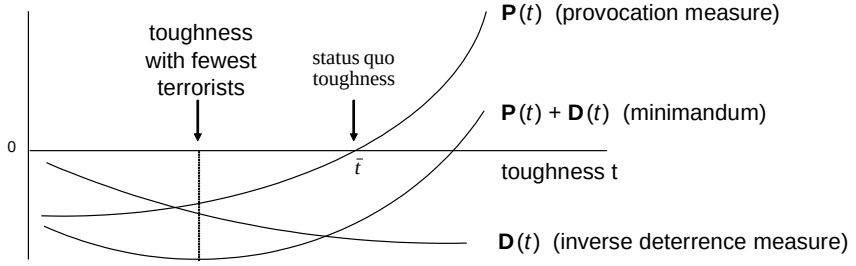


Figure 10: The trade-off in Theorem 2 for a unidimensional policy space $T$

Figure 10 illustrates the (one-dimensional) case that $t$ is a toughness level from a toughness interval $T \subseteq \mathbb{R}$. Plausibly, the higher toughness $t$, the larger $P(t)$ (more provocation) and the smaller $D(t)$ (more deterrence), the goal being to minimise the sum $P(t)+D(t)$. If $T$ is multidimensional, the trade-off becomes multi-dimensional, possibly with cause-related dimensions.

[21] $v' : \mathbb{R} \to \mathbb{R}$ is a strictly decreasing function, so has strictly decreasing inverse function $v'^{-1} : v(\mathbb{R}) \to \mathbb{R}$.

By contrast, the provocation-neglecting policy maker believes that $P(t) = 0$ (no provocation) for all (arbitrarily tough) policies $t \in T$, hence what he minimises is not $P(t) + D(t)$ but $D(t)$, leading to toughness exaggeration.

Perhaps surprisingly, the optimal policies in Theorem 2 do not depend on the distribution of types across the population: minimising $P(t) + D(t)$ is optimal regardless of how many individuals are highly damage-inclined (large $\theta \in \mathbb{R}$). So the policy maker can set its policy without 'understanding' people. This interesting feature of the model (with its stylised notion of optimality: minimising the *number* of terrorists) is certainly an exception. In other models, optimal policies are type-distribution-sensitive and often analytically intractable.

Finally, the provocation and deterrence measures $P(t)$ and $D(t)$ differ from the earlier-studied provocation and deterrence effects $\mathbf{PE}(t)$ and $\mathbf{DE}(t)$ because they arise in the context of minimising the *number* of terrorists, not the *sum-total amount* of terrorism as earlier.

# B Proof of Theorems 1 and 2

*Proof of Theorem 1.* Suppose UP and R1-R3.

(a) By $\mathrm{x}(t) = \mathrm{x}(t,t)$, the function $t \mapsto \mathrm{x}(t)$ is the composition of the differentiable functions $t \mapsto (t,t)$ (from $T$ to $T \times T$) and the by R3 differentiable function $(t_1, t_2) \mapsto \mathrm{x}(t_1, t_2)$ (from $T \times T$ to $\mathbb{R}_+$). Hence, by the chain rule, $t \mapsto \mathrm{x}(t)$ is itself differentiable and

$$\mathrm{x}'(t) = \frac{\partial}{\partial t_1}\mathrm{x}(t,t) + \frac{\partial}{\partial t_2}\mathrm{x}(t,t) \text{ at all } t \in T.$$

Setting $t = \bar{t}$, the left-hand side becomes $E$, and the right-hand side is recognised as the sum of $PE$ and $DE$; for instance,

$$\frac{\partial}{\partial t_1}\mathrm{x}(\bar{t},\bar{t}) = \left.\frac{\partial}{\partial t_1}\mathrm{x}(t,\bar{t})\right|_{t=\bar{t}} = \left.\frac{d}{dt}\mathrm{x}_{\mathrm{prov}}(t)\right|_{t=\bar{t}} = PE.$$

(b) Throughout, I write $(\bar{x}, \bar{y})$ for the pair $(\mathrm{x}(\bar{t},\bar{t}), \mathrm{y}(\bar{t},\bar{t}))$ $(= (\mathrm{x}_{\mathrm{prov}}(\bar{t}), \mathrm{y}_{\mathrm{prov}}(\bar{t})) = (\mathrm{x}_{\mathrm{deter}}(\bar{t}), \mathrm{y}_{\mathrm{deter}}(\bar{t})))$.

1. In this part I assume NIP and show that $PE \geq 0$. Suppose for a contradiction that $PE < 0$. I establish several claims; the last one contains the desired contradiction.

*Claim 1.* There exists a toughness level $\tilde{t} \in T$ larger than $\bar{t}$ such that $\mathrm{x}_{\mathrm{prov}}(t) < \bar{x}$ for all $t \in (\bar{t}, \tilde{t}]$. I write $\tilde{x} := \mathrm{x}_{\mathrm{prov}}(\tilde{t})$ and $\tilde{y} := \mathrm{y}_{\mathrm{prov}}(\tilde{t})$.

As $PE = \lim_{t\to\bar{t}} \frac{\mathrm{x}_{\mathrm{prov}}(t) - \mathrm{x}_{\mathrm{prov}}(\bar{t})}{t-\bar{t}} = \lim_{t\to\bar{t}} \frac{\mathrm{x}_{\mathrm{prov}}(t) - \bar{x}}{t-\bar{t}}$ and as $PE < 0$, we have $\frac{\mathrm{x}_{\mathrm{prov}}(t) - \bar{x}}{t-\bar{t}} < 0$ for all $t \neq \bar{t}$ in a sufficiently small neighbourhood of $\bar{t}$. In particular,

there is a $\tilde{t} > \bar{t}$ such that for all $t \in (\bar{t}, \tilde{t}]$ we have $\mathrm{x}_{\mathrm{prov}}(t) - \bar{x} < 0$, i.e. $\mathrm{x}_{\mathrm{prov}}(t) < \bar{x}$, q.e.d.

*Claim 2.* For every damage level $x \in [\tilde{x}, \bar{x}]$ there exists a toughness level $t \in [\bar{t}, \tilde{t}]$ such that $\mathrm{x}_{\mathrm{prov}}(t) = x$.

Let $x \in [\tilde{x}, \bar{x}]$. As the function $\mathrm{x}_{\mathrm{prov}}$ is continuous on $[\bar{t}, \tilde{t}]$ (because it is differentiable) and as $\mathrm{x}_{\mathrm{prov}}(\tilde{t}) = \tilde{x} \leq x \leq \bar{x} = \mathrm{x}_{\mathrm{prov}}(\bar{t})$, the intermediate value theorem implies the existence of a $t \in [\bar{t}, \tilde{t}]$ such that $\mathrm{x}_{\mathrm{prov}}(t) = x$, q.e.d.

For every damage level $x \in [\tilde{x}, \bar{x}]$, define

$$S(x) := \sup\{y \geq 0 : u_{\bar{t}}(x, y) \geq u_{\bar{t}}(\tilde{x}, \tilde{y})\} \ (\in \mathbb{R}_+ \cup \{\infty, -\infty\}),$$

with the usual conventions that $\sup \emptyset := -\infty$ and that $\sup Q := \infty$ whenever $Q \subseteq \mathbb{R}$ has no upper bound.

*Claim 3.* $S(\tilde{x}) = \tilde{y}$.

By definition of $S(\tilde{x})$ we have $S(\tilde{x}) \geq \tilde{y}$, and using punishment-aversion it follows that $S(\tilde{x}) = \tilde{y}$, q.e.d.

*Claim 4.* For all $x \in [\tilde{x}, \bar{x}]$, if $S(x) \in \mathbb{R}_+$ then $u_{\bar{t}}(x, S(x)) = u_{\bar{t}}(\tilde{x}, \tilde{y})$.

Consider any $x \in [\tilde{x}, \bar{x}]$ with $S(x) \in \mathbb{R}_+$. To show that $u_{\bar{t}}(x, S(x)) \geq u_{\bar{t}}(\tilde{x}, \tilde{y})$, note that by definition of $S(x)$ there is a sequence $(y_k)_{k=1,2,...}$ in $[0, S(s)]$ converging to $S(x)$ such that $u_{\bar{t}}(x, y_k) \geq u_{\bar{t}}(\tilde{x}, \tilde{y})$ for all $k = 1, 2, ...$ As $u_{\bar{t}}$ is a continuous function, $u_{\bar{t}}(x, y_k) \to u_{\bar{t}}(x, S(x))$ as $k \to \infty$. So, as weak inequalities are preserved in the limit, $u_{\bar{t}}(x, S(x)) \geq u_{\bar{t}}(\tilde{x}, \tilde{y})$. To show the converse inequality, consider any sequence $(z_k)_{k=1,2,...}$ in $(S(x), \infty)$ converging to $S(x)$ (of course there is one). By definition of $S(x)$, we have $u_{\bar{t}}(x, z_k) < u_{\bar{t}}(\tilde{x}, \tilde{y})$ for all $k = 1, 2, ...$ So, again by continuity of $u_{\bar{t}}$, $u_{\bar{t}}(x, S(x)) \leq u_{\bar{t}}(\tilde{x}, \tilde{y})$, q.e.d.

*Claim 5.* $S(\bar{x}) = \infty$.

For a contradiction, suppose $S(\bar{x}) \neq \infty$. We also have $S(\bar{x}) \neq -\infty$ because $u_{\bar{t}}(\bar{x}, \bar{y}) \geq u_{\bar{t}}(\tilde{x}, \tilde{y})$ (as $(\bar{x}, \bar{y})$ maximises $u_{\bar{t}}(x, y)$ subject to $(x, y) \in \mathbf{F}_{\bar{t}}$). So $S(\bar{x}) \in \mathbb{R}_+$. Hence, by Claim 4,

(*) $u_{\bar{t}}(\bar{x}, S(\bar{x})) = u_{\bar{t}}(\tilde{x}, \tilde{y})$.

This and the inequality $u_{\bar{t}}(\bar{x}, \bar{y}) > u_{\bar{t}}(\tilde{x}, \tilde{y})$ (which holds because $(\bar{x}, \bar{y})$ uniquely maximises $u_{\bar{t}}(x, y)$ subject to $(x, y) \in \mathbf{F}_{\bar{t}}$) imply that $u_{\bar{t}}(\bar{x}, \bar{y}) > u_{\bar{t}}(\bar{x}, S(\bar{x}))$, which by punishment-aversion entails that

(**) $\bar{y} < S(\bar{x})$.

But (*) also implies that $u_{\tilde{t}}(\tilde{x}, \tilde{y}) \leq u_{\tilde{t}}(\bar{x}, S(\bar{x}))$ by NIP. Using (**) and punishment-aversion, it follows that $u_{\tilde{t}}(\tilde{x}, \tilde{y}) < u_{\tilde{t}}(\bar{x}, \bar{y})$, a contradiction since $(\tilde{x}, \tilde{y})$ maximises $u_{\tilde{t}}(x, y)$ subject to $(x, y) \in \mathbf{F}_{\bar{t}}$, q.e.d.

*Claim 6.* There exists an $x \in [\tilde{x}, \bar{x}]$ that is smallest subject to $S(x) = \infty$. I denote it by $x_\infty$.

I have to show that the set $X_\infty := \{x \in [\tilde{x}, \bar{x}] : S(x) = \infty\}$ has a smallest element. By Claim 5, $X_\infty$ is non-empty. So it has an infimum $x_* := \inf X_\infty$ in $[\tilde{x}, \bar{x}]$. I have to show that $S(x_*) = \infty$ (i.e. that the infimum is a minimum). By definition of $x_*$, there is a sequence $(x_k)_{k=1,2,...}$ in $X_\infty$ that converges to $x_*$. Consider any $y \geq 0$. As $u_{\bar{t}}$ is a continuous function, $u_{\bar{t}}(x_k, y) \to u_{\bar{t}}(x_*, y)$ as $k \to \infty$. Note also that, for all $k$, $u_{\bar{t}}(x_k, y) \geq u_{\bar{t}}(\tilde{x}, \tilde{y})$: otherwise $u_{\bar{t}}(x_k, y') < u_{\bar{t}}(\tilde{x}, \tilde{y})$ for all $y' > y$ by punishment-aversion, implying that $S(x_k) \leq y$, in contradiction with $x_k \in X_\infty$. So, as weak inequalities are preserved in the limit, we have $u_{\bar{t}}(x_*, y) \geq u_{\bar{t}}(\tilde{x}, \tilde{y})$. Since this has been shown for all $y \geq 0$, we have $S(x_*) = \infty$, q.e.d.

*Claim 7.* $x_\infty > \tilde{x}$.

By Claim 3, $S(\tilde{x}) = \tilde{y}$. So $S(\tilde{x}) < \infty$. Hence $x_\infty \neq \tilde{x}$ by Claim 6, q.e.d.

*Claim 8.* $S(x) \to \infty$ as $x \uparrow x_\infty$.

Consider any sequence $(x_k)_{k=1,2,...}$ in $[\tilde{x}, x_\infty)$ such that $x_k \uparrow x_\infty$. I have to show that $S(x_k) \to \infty$. For a contradiction, suppose that $S(x_k) \not\to \infty$. Then there is a $\hat{y} > 0$ and a subsequence $(x_{k_j})_{j=1,2,...}$ – I denote it simply by $(x'_j)_{j=1,2,...}$ – such that $S(x'_j) < \hat{y}$ for all $j$. So, by definition of $S(x'_j)$, we have $u_{\bar{t}}(x'_j, \hat{y}) < u_{\bar{t}}(\tilde{x}, \tilde{y})$ for all $j$. Hence, as $u_{\bar{t}}(x'_j, \hat{y}) \to u_{\bar{t}}(x_\infty, \hat{y})$ by continuity of $u_{\bar{t}}$ and as weak inequalities are preserved in the limit, we have $u_{\bar{t}}(x_\infty, \hat{y}) \leq u_{\bar{t}}(\tilde{x}, \tilde{y})$. Hence, by punishment-aversion, we have $u_{\bar{t}}(x_\infty, y) < u_{\bar{t}}(\tilde{x}, \tilde{y})$ for all $y > \hat{y}$. So $S(x_\infty) \leq \hat{y}$, a contradiction since $S(x_\infty) = \infty$ by Claim 6, q.e.d.

*Claim 9.* For all $y \geq 0$, we have $u_{\tilde{t}}(x_\infty, y) \geq u_{\tilde{t}}(\tilde{x}, \tilde{y})$.

Let $y \geq 0$. By Claim 8 there is an $x' \in [\tilde{x}, x_\infty)$ such that $S(x) \geq y$ for all $x \in (x', x_\infty)$. For any $x \in (x', x_\infty)$, we have

- $u_{\tilde{t}}(x, y) \geq u_{\tilde{t}}(x, S(x))$ by punishment-aversion, and
- $u_{\tilde{t}}(x, S(x)) \geq u_{\tilde{t}}(\tilde{x}, \tilde{y})$ by NIP and the fact that $u_{\bar{t}}(x, S(x)) = u_{\bar{t}}(\tilde{x}, \tilde{y})$ given Claim 4.

So $u_{\tilde{t}}(x, y) \geq u_{\tilde{t}}(\tilde{x}, \tilde{y})$ for all $x \in (x', x_\infty)$. Hence, as $u_{\tilde{t}}$ is continuous, $u_{\tilde{t}}(x_\infty, y) \geq u_{\tilde{t}}(\tilde{x}, \tilde{y})$, q.e.d.

*Claim 10.* $(\tilde{x}, \tilde{y})$ does not maximise $u_{\tilde{t}}(x, y)$ subject to $(x, y) \in \mathbf{F}_{\bar{t}}$ (a contradiction, completing the proof).

By Claim 2 there is a toughness level $t \in [\bar{t}, \tilde{t}]$ such that $\mathrm{x}_{\mathrm{prov}}(t) = x_\infty$. Write $y_\infty := \mathrm{y}_{\mathrm{prov}}(t)$. I now apply Claim 9 to $y_\infty + 1$, which gives us $u_{\tilde{t}}(x_\infty, y_\infty + 1) \geq u_{\tilde{t}}(\tilde{x}, \tilde{y})$. So, $u_{\tilde{t}}(x_\infty, y_\infty) > u_{\tilde{t}}(\tilde{x}, \tilde{y})$ by punishment-aversion. As $(x_\infty, y_\infty) = (\mathrm{x}_{\mathrm{prov}}(t), \mathrm{y}_{\mathrm{prov}}(t)) \in \mathbf{F}_{\bar{t}}$, it follows that $(\tilde{x}, \tilde{y})$ does not maximise $u_{\tilde{t}}(x, y)$ subject to $(x, y) \in \mathbf{F}_{\bar{t}}$, q.e.d.

2. I now assume TC and MP, and show that $DE \leq 0$. Suppose that $DE > 0$. I derive a contradiction, again in several steps.

*Claim 1.* The feasible damage sets $\mathbf{X}_t \subseteq \mathbb{R}_+$, $t \in T$, are intervals containing 0, and the function $t \mapsto \mathbf{X}_t$ is weakly decreasing (w.r.t. set-inclusion).

Each $\mathbf{X}_t$ is an interval containing 0 because it is the projection on the $x$-coordinate of the set $\mathbf{F}_t \subseteq \mathbb{R}^2_+$, which by MP is connected and contains $(0,0)$. Now consider $t < t'$ in $T$. To show that $\mathbf{X}_{t'} \subseteq \mathbf{X}_t$, let $x \in \mathbf{X}_{t'}$. By the increasingness assumption in MP, $f'_t(x) \leq f'_{t'}(x)$; so $f'_t(x) < \infty$, and hence $x \in \mathbf{X}_{t'}$, q.e.d.

*Claim 2.* There exists a toughness level $\tilde{t} \in T$ larger than $\bar{t}$ such that $\mathrm{x}_{\mathrm{deter}}(\tilde{t}) > \bar{x}$. I henceforth write $\tilde{x}$ for $\mathrm{x}_{\mathrm{deter}}(\tilde{t})$.

By assumption $DE = \mathrm{x}'_{\mathrm{deter}}(\bar{t}) > 0$, i.e. $\lim_{t\to\bar{t}} \frac{\mathrm{x}_{\mathrm{deter}}(t)-\mathrm{x}_{\mathrm{deter}}(\bar{t})}{t-\bar{t}} = \lim_{t\to\bar{t}} \frac{\mathrm{x}_{\mathrm{deter}}(t)-\bar{x}}{t-\bar{t}} > 0$. So $\frac{\mathrm{x}_{\mathrm{deter}}(t)-\bar{x}}{t-\bar{t}} > 0$ for all $t \neq \bar{t}$ sufficiently close $\bar{t}$. Hence $\mathrm{x}_{\mathrm{deter}}(t) - \bar{x} > 0$ for all $t > \bar{t}$ sufficiently close $\bar{t}$, q.e.d.

*Claim 3.* $f_t(x)$ is (at least weakly) increasing in each argument, i.e. in $x \in \mathbf{X}_t$ (for each $t \in T$) and in $t \in \{t' \in T : x \in \mathbf{X}_{t'}\}$ (for each $x \geq 0$).

At any fixed $t \in T$, increasingness in $x \in \mathbf{X}_t$ holds since, by MP, $f'_t(x) \geq 0$ at all $x \in \mathbf{X}_t$. To show increasingness in $t \in T$, consider a fixed $\hat{x} \geq 0$, and let $t_-, t_+ \in \{t \in T : \hat{x} \in \mathbf{X}_t\}$ satisfy $t_- \leq t_+$. To show that $f_{t_-}(\hat{x}) \leq f_{t_+}(\hat{x})$, I define the function $g : [0, \hat{x}] \to \mathbb{R}$ by $g(x) := f_{t_+}(x) - f_{t_-}(x)$ and show that $g(\hat{x}) \geq 0$. This follows from two facts:

- $g(\hat{x}) \geq g(0)$, by the following argument. At every $x \in [0, \hat{x}]$, $g$ has a non-negative derivative $g'(x) \geq 0$ because $t_+ \geq t_-$ and because $f'_t(x)$ is by MP increasing in $t \in T$. So $g$ is increasing, implying that $g(\hat{x}) \geq g(0)$.
- $g(0) = 0$, by the following argument. For all $t \in T$ we have $f_t(0) = 0$ because $(0,0) \in \mathbf{F}_t$ by MP. So $g(0) = f_{t_+}(0) - f_{t_-}(0) = 0 - 0 = 0$, q.e.d.

*Claim 4.* For all $t \in T$, $\mathrm{y}_{\mathrm{deter}}(t) = f_t(\mathrm{x}_{\mathrm{deter}}(t))$ (so I may henceforth write $f_{\bar{t}}(\bar{x})$ for $\mathrm{y}_{\mathrm{deter}}(\bar{t})$, and $f_{\tilde{t}}(\tilde{x})$ for $\mathrm{y}_{\mathrm{deter}}(\tilde{t})$).

Let $t \in T$. As $\mathbf{F}_t$ is topologically closed, it contains $(\mathrm{x}_{\mathrm{deter}}(t), f_t(\mathrm{x}_{\mathrm{deter}}(t)))$ (by definition of $f_t$), i.e. $(\mathrm{x}_{\mathrm{deter}}(t), f_t(\mathrm{x}_{\mathrm{deter}}(t)))$ is a feasible choice. So $\mathrm{y}_{\mathrm{deter}}(t)$ cannot be larger than $f_t(\mathrm{x}_{\mathrm{deter}}(t))$ (otherwise $(\mathrm{x}_{\mathrm{deter}}(t), \mathrm{y}_{\mathrm{deter}}(t))$ would be dispreferred to $(\mathrm{x}_{\mathrm{deter}}(t), f_t(\mathrm{x}_{\mathrm{deter}}(t)))$ by punishment-aversion); it also cannot it be smaller than $f_t(\mathrm{x}_{\mathrm{deter}}(t))$ (otherwise $(\mathrm{x}_{\mathrm{deter}}(t), \mathrm{y}_{\mathrm{deter}}(t))$ would be infeasible by definition of $f_t$), q.e.d.

*Claim 5.* $\bar{x}, \tilde{x} \in \mathbf{X}_{\tilde{t}} \subseteq \mathbf{X}_{\bar{t}}$ and $f_{\tilde{t}}(\tilde{x}) - f_{\tilde{t}}(\bar{x}) \geq f_{\bar{t}}(\tilde{x}) - f_{\bar{t}}(\bar{x}) \geq 0$.

By definition, $\tilde{x} \in \mathbf{X}_{\tilde{t}}$. Also $\bar{x} \in \mathbf{X}_{\tilde{t}}$, as $0 \leq \bar{x} < \tilde{x}$ and as $\mathbf{X}_{\tilde{t}}$ is an interval containing 0 (by Claim 1) and containing $\tilde{x}$. By Claim 1, $\mathbf{X}_{\tilde{t}} \subseteq \mathbf{X}_{\bar{t}}$. As for the inequalities, the second one holds because $\tilde{x} > \bar{x}$ and because $f_{\bar{t}}(x)$ is increasing in $x \in \mathbf{X}_{\bar{t}}$ by Claim 3. I now show the first inequality. For every $x \in \mathbf{X}_{\tilde{t}}$, since $f'_t(x)$ is by MP increasing in $t \in T$, and since $\tilde{t} > \bar{t}$, we have $f'_{\tilde{t}}(x) \geq f'_{\bar{t}}(x)$,

i.e. $\frac{d}{dx}(f_{\tilde{t}}(x) - f_{\bar{t}}(x)) \geq 0$. So $f_{\tilde{t}}(x) - f_{\bar{t}}(x)$ is a weakly increasing function of $x \in \mathbf{X}_{\tilde{t}}$. Hence, as $\tilde{x} > \bar{x}$, we have $f_{\tilde{t}}(\tilde{x}) - f_{\bar{t}}(\tilde{x}) \geq f_{\tilde{t}}(\bar{x}) - f_{\bar{t}}(\bar{x})$, or by reordering, $f_{\tilde{t}}(\tilde{x}) - f_{\tilde{t}}(\bar{x}) \geq f_{\bar{t}}(\tilde{x}) - f_{\bar{t}}(\bar{x})$, q.e.d.

*Claim 6.* There is a counterexample to the condition TC.

I consider any $\delta > 0$ and show that TC is violated when applied to the pairs $(\bar{x}, f_{\bar{t}}(\bar{x}))$ and $(\tilde{x}, f_{\bar{t}}(\tilde{x}) + \delta)$ and to the 'punishment shift' $\epsilon := f_{\tilde{t}}(\bar{x}) - f_{\bar{t}}(\bar{x})$. To see that TC is applicable here, note three things.

- The first of the two pairs has lower punishment because $f_{\bar{t}}(\bar{x}) \leq f_{\bar{t}}(\tilde{x}) < f_{\bar{t}}(\tilde{x}) + \delta$, where the first inequality follows from Claim 5.
- The first pair is preferred to the second pair because $(\bar{x}, f_{\bar{t}}(\bar{x})) \succ_{\bar{t}} (\tilde{x}, f_{\bar{t}}(\tilde{x})) \succ_{\bar{t}} (\tilde{x}, f_{\bar{t}}(\tilde{x}) + \delta)$, where the second preference holds by punishment-aversion, and the first by the fact that $(\bar{x}, f_{\bar{t}}(\bar{x}))$ uniquely maximises $\succeq_{\bar{t}}$ within $\mathbf{F}_{\bar{t}}$.
- $\epsilon \geq 0$ because $\tilde{t} > \bar{t}$ and because $f_t(\bar{x})$ is by Claim 3 increasing in $t$. (To be precise, the condition TC is originally stated with a strict inequality '$\epsilon > 0$', but the condition trivially holds also if $\epsilon = 0$.)

Now, by TC, $(\bar{x}, f_{\bar{t}}(\bar{x}) + \epsilon) \succeq_{\bar{t}} (\tilde{x}, f_{\bar{t}}(\tilde{x}) + \delta + \epsilon)$. So, as $\epsilon = f_{\tilde{t}}(\bar{x}) - f_{\bar{t}}(\bar{x})$,

$$(\bar{x}, f_{\bar{t}}(\bar{x})) \succeq_{\bar{t}} (\tilde{x}, f_{\bar{t}}(\tilde{x}) + f_{\tilde{t}}(\bar{x}) - f_{\bar{t}}(\bar{x}) + \delta).$$

In this, the right-hand side is weakly preferred to $(\tilde{x}, f_{\tilde{t}}(\tilde{x}) + \delta)$, by punishment-aversion as by Claim 5 $f_{\bar{t}}(\tilde{x}) + f_{\tilde{t}}(\bar{x}) - f_{\bar{t}}(\bar{x}) + \delta \leq f_{\tilde{t}}(\tilde{x}) + \delta$. So

$$(\bar{x}, f_{\bar{t}}(\bar{x})) \succeq_{\bar{t}} (\tilde{x}, f_{\tilde{t}}(\tilde{x}) + \delta).$$

Since we have shown this for *every* $\delta > 0$, and since by the continuity of $\succeq_{\bar{t}}$ any weak preference is preserved in the limit, we deduce that

$$(\bar{x}, f_{\bar{t}}(\bar{x})) \succeq_{\bar{t}} \lim_{\delta \downarrow 0} (\tilde{x}, f_{\tilde{t}}(\tilde{x}) + \delta) = (\tilde{x}, f_{\tilde{t}}(\tilde{x})),$$

a contradiction because $(\tilde{x}, f_{\tilde{t}}(\tilde{x}))$ uniquely maximises $\succeq_{\tilde{t}}$ within $\mathbf{F}_{\tilde{t}}$ and because $(\bar{x}, f_{\tilde{t}}(\bar{x})) \in \mathbf{F}_{\tilde{t}}$ (by definition of $f_{\tilde{t}}$ and the topological closedness of $\mathbf{F}_{\tilde{t}}$). ■

*Proof of Theorem 2.* Assume PM and CP. A close look at part (c) reveals that it follows from part (b) by taking the case that $P(t) = 0$ for all $t \in T$ (no policy provokes); similarly, in (d), the claim referring to (c) follows from that referring to (b) by taking the mentioned special case. The proof of (a), (b) and the relevant part of (d) is done in several claims; (a) follows from Claims 2 and 9, (b) from Claim 10, and the relevant part of (d) from Claim 5.

*Claim 1.* For every policy $t \in T$, the feasible damage set $\mathbf{X}_t \subseteq \mathbb{R}_+$ is a non-singleton interval containing 0, and the ratio $f_t(x)/x$ is an increasing function of $x \in \mathbf{X}_t \backslash \{0\}$.

Let $t \in T$. As $\mathbf{F}_t$ is by CP connected and contains $(0,0)$, $\mathbf{X}_t$ is an interval containing 0. $\mathbf{X}_t$ is non-singleton: otherwise $\mathbf{F}_t = \{(0,0)\}$, so that $(0,0)$ would be each type's best response to policy $t$, a trivial case excluded in PM. Now let $x, x' \in \mathbf{X}_t \backslash \{0\}$ with $x < x'$. The $x$ can be written as $\lambda x'$ for some $0 < \lambda < 1$. So, as $f_t$ is by CP convex, $f(x) \leq \lambda f(x') + (1-\lambda) f(0) = \lambda f(x')$, whence $f(x)/x \leq \lambda f(x')/x = f(x')/x'$, q.e.d.

*Claim 2.* For every policy $t \in T$, the function $f_t : \mathbf{X}_t \to \mathbb{R}_+$ is increasing, continuous, and differentiable (from the right) at $x = 0$ with $f_t'(0) \geq 0$.

Let $t \in T$. The function $f_t$ is increasing on $\mathbf{X}_t \backslash \{0\}$ by Claim 1, hence increasing on its full domain as $f_t(0) = 0$ by $(0,0) \in \mathbf{F}_t$. As $x \downarrow 0$, the ratio $\frac{f(x)-f(0)}{x-0} = f(x)/x$ is decreasing (by Claim 1) and bounded below by 0, hence has a limit $f'(0)$ that is moreover $\geq 0$. It remains to show continuity. Since $f_t$ is convex, it is continuous on every open subinterval $I \subseteq \mathbf{X}_t$ (as the reader can easily check); so the only potential discontinuities arise at the boundaries of $\mathbf{X}_t$. Continuity at the left boundary 0 holds by differentiability. Now suppose $\mathbf{X}_t$ contains a right boundary, i.e. takes the form $\mathbf{X}_t = [0, x^*]$. and consider an increasing sequence $(x_k)_{k=1,2,...}$ in $\mathbf{X}_t$ with $x_k \uparrow x^*$. As $f_t$ is increasing, the sequence $(f_t(x_k))_{k=1,2,...}$ is increasing and bounded (by $f_t(x^*)$), hence converges to a value $y$. So $(x_k, f_t(x_k))_{k=1,2,...}$ is a sequence in $\mathbf{F}_t$ that converges to $(a^*, y)$. So, as $\mathbf{F}_t$ is topologically closed, we have $(a^*, y) \in \mathbf{F}_t$, hence $f(a^*) = y$, q.e.d.

*Claim 3.* The function $d : \mathbb{R}_+ \to \mathbb{R}$ is strictly increasing.

This follows easily from punishment-aversion, q.e.d.

*Claim 4.* For every policy $t \in T$, if $\mathbf{X}_t = [0, a^*)$ for some $a^* := a_t^* \in \mathbb{R}_+$, then $f_t(x) \to \infty$ as $x \uparrow a^*$, and for every type $\theta$ there exists an $\epsilon = \epsilon_\theta \in (0, a^*)$ such that $(x, f_t(x)) \prec_{t,\theta} (0,0)$ for all $x \in (a^* - \epsilon, a^*)$.

Let $t \in T$ and $\mathbf{X}_t = [0, a^*)$ with $a^* \in \mathbb{R}_+$.

First, if we had $f_t(x) \not\to \infty$ as $x \uparrow a^*$, there would exist a sequence $(x_k)_{k=1,2,...}$ in $\mathbf{X}_t$ with $x_k \to a^*$ such that the sequence $(x_k, f_t(x_k))_{k=1,2,...}$ is bounded in $\mathbb{R}_+^2$; now taking any convergent subsequence of it (there is one; see Th. A1.8 in Geoffrey Jehle and Philip Reny 2001), we would have a sequence in $\mathbf{F}_t$ that converges to some point $(a^*, y)$ outside $\mathbf{F}_t$, a contradiction since $\mathbf{F}_t$ is topologically closed.

Second, consider any type $\theta$. The claim is obvious if $\theta$ is a peaceful type. Now suppose $\theta \in \mathbb{R}$ and for a contradiction let no $\epsilon > 0$ have the required property. Then there exists a sequence $(x_k)_{k=1,2,...}$ in $\mathbf{X}_t$ such that $x_k \to a^*$ and $(x_k, f_t(x_k)) \succeq_{t,\theta} (0,0)$ for all $k = 1, 2, ...$ By $f_t(x_k) \to \infty$ and $x_k \to a^*$, we have $\alpha_k := f_t(x_k)/x_k \to \infty$ as $k \to \infty$. In other words, the slope $\alpha_k$ of the straight line from $(0,0)$ to $(x_k, f_t(x_k))$ tends to $\infty$ as $k \to \infty$. The point $(1/\alpha_k, 1)$ is on this line; so, as $(x_k, f_t(x_k)) \succeq_{t,\theta} (0,0)$ and as preference $\succeq_{t,\theta}$ is convex, we have

$(1/\alpha_k, 1) \succeq_{t,\theta} (0,0)$ for all $k = 1, 2, ...$ It follows that $(0,1) \succeq_{t,\theta} (0,0)$, using that $(1/\alpha_k, 1) \to (0,1)$ and that weak preferences are preserved in the limit because $\succeq_{t,\theta}$ is continuous (by the continuity of $u_{t,\theta} : \mathbb{R}^2_+ \to \mathbb{R}$). This is a contradiction, because punishment-aversion requires that $(0,1) \prec_{t,\theta} (0,0)$, q.e.d.

*Claim 5.* Each type $\theta$ has a unique optimal response to each $t \in T$, denoted $(x_t^\theta, y_t^\theta)$; this optimum satisfies $y_t^\theta = f_t(x_t^\theta)$, and it is $(0,0)$ if $\theta$ is a peaceful type or a type in $\mathbb{R}$ with $\theta + P(t) \leq 0$.

Consider a $t \in T$. If a type has an optimum $(x, y) \in \mathbf{F}_t$ then it must be that $y = f_t(x)$, by punishment-aversion and since $\mathbf{F}_t$ contains $(x, f_t(x))$ by topological closedness (see CP).

All peaceful types $\theta$ and also all types $\theta \in \mathbb{R}$ with $\theta + P(t) \leq 0$ have peaceful (and punishment-averse) preferences $\succeq_{t,\theta}$, hence have the no-damage-no-punishment pair $(0,0)$ as their unique optimal response to $t$.

It remains to consider a type $\theta \in \mathbb{R}$ with $\theta + P(t) > 0$.

To show uniqueness, suppose for a contradiction that $(x_1, f_t(x_1))$ and $(x_2, f_t(x_2))$ are two distinct optimal responses to $t$. Consider any $x^*$ strictly between $x_1$ and $x_2$, say $x^* = \lambda x_1 + (1-\lambda)x_2$ where $\lambda \in (0,1)$. The response $(x^*, f_t(x^*))$ is of course feasible (i.e. in $\mathbf{F}_t$); I show that it is strictly preferred to $(x_1, f_t(x_1))$, a contradiction. Let $y^* := \lambda f_t(x_1) + (1-\lambda) f_t(x_2)$. As $f_t$ is convex, $y^* \geq f_t(x^*)$, and so by punishment-aversion, (*) $(x^*, f_t(x^*)) \succeq_{t,\theta} (x^*, y^*)$. Moreover, by $(x^*, y^*) = \lambda(x_1, f_t(x_1)) + (1-\lambda)(x_2, f_t(x_2))$, the strict convexity of $\succeq_{t,\theta}$ implies that $(x^*, y^* + \epsilon) \succeq_{t,\theta} (x_1, f_t(x_1))$ for some $\epsilon > 0$, and hence that $(x^*, y^*) \succ_{t,\theta} (x_1, f_t(x_1))$ by punishment-aversion. The latter and (*) imply that $(x^*, f_t(x^*)) \succ_{t,\theta} (x_1, f_t(x_1))$, the desired contradiction.

To show existence, I distinguish three cases.

*Case 1*: $\mathbf{X}_t$ is bounded and closed, i.e. of the compact form $\mathbf{X}_t = [0, x_t^*]$. As $f_t$ is continuous (by Claim 2) and $u_{t,\theta}$ is also continuous, $u_{t,\theta}(x, f_t(x))$ is a continuous function of $x \in \mathbf{X}_t$, hence (as $\mathbf{X}_t$ is compact) admits a global maximum $\tilde{x} \in \mathbf{X}_t$ by Weierstrass' Theorem (Carl Simon and Lawrence Blume 1994, Th. 30.1). Now $(\tilde{x}, f_t(\tilde{x}))$ maximises $u_{t,\theta}(x, y)$ subject to $(x, y) \in \mathbf{F}_t$, as desired.

*Case 2*: $\mathbf{X}_t$ is bounded and open, i.e. of the form $\mathbf{X}_t = [0, x_t^*)$. By Claim 4, $\mathbf{X}_t$ has a subinterval of the form $[0, x_t^* - \epsilon]$ such that all feasible pairs $(x, f_t(x))$ with $x \notin [0, x_t^* - \epsilon]$ are strictly dispreferred to some feasible pair $(x, f_t(x))$ with $x \in [0, x_t^* - \epsilon]$ (in fact, strictly dispreferred to $(0,0)$). So it suffices to show that the function $u_{t,\theta}(x, f_t(x))$ of $x$ admits a maximum on the compact subinterval $[0, x_t^* - \epsilon] \subseteq \mathbf{X}_t$. This can be done analogously to the proof in case 1.

*Case 3*: $\mathbf{X}_t$ is unbounded, i.e. $\mathbf{X}_t = \mathbb{R}_+$. By an argument analogous to that

in case 1, on the compact subinterval $[0, \theta + P(t)]$ the function $u_{t,\theta}(x, f_t(x))$ of $x$ admits a maximum $\tilde{x} \in [0, \theta + P(t)]$. I show that $\tilde{x}$ is a global maximum of $u_{t,\theta}(x, f_t(x))$, which completes the proof as it establishes that $(\tilde{x}, f_t(\tilde{x}))$ is an optimal within $\mathbf{F}_t$. To do so, consider any $x \in \mathbb{R}_+$ larger than $\theta + P(t)$. Recall that

$$u_{t,\theta}(x, f_t(x)) = v(x - \theta - P(t)) - d(f_t(x)).$$

So, as $v(x-\theta-P(t)) \leq v(0)$ (because $v$ peaks at 0) and $d(f_t(x)) \geq d(f_t(\theta+P(t)))$ (because $f_t$ and $d$ are increasing functions by Claims 2 and 3),

$$u_{t,\theta}(x, f_t(x)) \leq v(0) - d(f_t(\theta + P(t))) = u_{t,\theta}(\theta + P(t)), f_t(\theta + P(t))),$$

which is at most $u_{t,\theta}(\tilde{x}, f_t(\tilde{x}))$ by the restricted maximality property of $\tilde{x}$. This shows the global maximality property of $\tilde{x}$, q.e.d.

*Claim 6.* The derivative $v' : \mathbb{R} \to \mathbb{R}$ is strictly decreasing and $v'(0) = 0$.

This holds because $v$ is strictly concave and peaks at $x = 0$.

*Claim 7.* The range $v'(\mathbb{R})$ of $v'$ is an interval containing 0.

By the previous claim, $0 = v'(0) \in v'(\mathbb{R})$. To prove that $v'(\mathbb{R})$ is an interval (which if $v'$ is continuous follows easily from the intermediate value theorem), it suffices to show that, for all $x_1, x_2 \in v'(\mathbb{R})$ with $x_1 < x_2$ and all $s$ between $v'(x_1)$ and $v'(x_2)$ there exists an $\tilde{x} \in [x_1, x_2]$ such that $v'(\tilde{x}) = s$. Consider such $x_1, x_2, s$. By the previous claim, $v'(x_2) < v'(x_1)$, and hence $v'(x_2) < s < v'(x_1)$. I distinguish two cases, and write $k$ for the coefficient $\frac{v(x_2)-v(x_1)}{x_2-x_1}$.

*Case 1*: $s \geq k$. The (differentiable) function $g(x) := v(x)-[v(x_1)+s(x-x_1)]$ of $x \in \mathbb{R}$ satisfies

$$g'(x_1) = v'(x_1) - s > 0.$$

So, as $g(x_1) = 0$, we have $g(\bar{x}) > 0$ for some $\bar{x} \in (x_1, x_2)$. Moreover,

$$\begin{aligned} g(x_2) &= v(x_2) - v(x_1) - s(x_2 - x_1) \\ &\leq v(x_2) - v(x_1) - k(x_2 - x_1) = 0 \end{aligned}$$

By $g(\bar{x}) > 0$ and $g(x_2) \leq 0$, and since $g$ is continuous, the intermediate value theorem implies the existence of some $\tilde{x} \in (\bar{x}, x_2]$ such that $g(\tilde{x}) = 0$. This means that $v(\tilde{x}) = v(x_1) + s(\tilde{x} - x_1)$, i.e. that $\frac{v(\tilde{x})-v(x_1)}{\tilde{x}-x_1} = s$. Now we can apply the mean value theorem to the differentiable function $v$, which guarantees the existence of an $x \in [x_1, \tilde{x}]$ such that $v'(x) = \frac{v(\tilde{x})-v(x_1)}{\tilde{x}-x_1} = s$, as desired.

*Case 2*: $s \leq k$. An argument analogous to that in Case 1 (based now on the function $g(x) := v(x) - [v(x_2) + s(x - x_2)]$ of $x \in \mathbb{R}$) guarantees again the existence of the desired $\tilde{x}$, q.e.d.

*Claim 8.* For all $t \in T$ and all $\theta \in \mathbb{R}$, (i) the function $U_{t,\theta}(x) := u_{t,\theta}(x, f_t(x))$ of $x \in \mathbf{X}_t$ is differentiable (from the right) at $x = 0$ with $U'_{t,\theta}(0) = v'(-\theta - P(t)) - d'(0)f'_t(0)$, and (ii) $x^\theta_t > 0 \Leftrightarrow U'_{t,\theta}(0) > 0 \Leftrightarrow v'(-\theta - P(t)) > d'(0)f'_t(0)$.

Let $t \in T$ and $\theta \in \mathbb{R}$. Note that $U_{t,\theta}(x) = v(x - \theta - P(t)) - d(f_t(x)))$ for all $x \in \mathbf{X}_t$. As $f_t$ is differentiable (from the right) at $x = 0$ by Claim 1, and as the functions $v$ and $d$ are differentiable, the chain rule implies that $U_{t,\theta}(x)$ is differentiable (from the right) at $x = 0$ with derivative given by

$$U'_{t,\theta}(0) = v'(-\theta - P(t)) - d'(f_t(0))f'_t(0) = v'(-\theta - P(t)) - d'(0)f'_t(0).$$

In (ii), the second equivalence is obvious by (i). Regarding the first equivalence, if $U'_{t,\theta}(0) > 0$ then, for some $x \in \mathbf{X}_t$, we have $U_{t,\theta}(0) < U_{t,\theta}(x)$, implying that $(0, 0) = (0, f_t(0)) \prec_{t,\theta} (x, f_t(x))$, hence that $(0, 0)$ is not optimal, and so $x_{t,\theta} > 0$. Conversely, suppose now that $x^\theta_t > 0$. To show $U'_{t,\theta}(0) > 0$, I consider two cases.

*Case 1*: $d'(0) = 0$. Then $U'(0) = v'(-\theta - P(t))$, which is positive because $-\theta - P(t) < 0$ (by $x^\theta_t > 0$ and Claim 5) and because $v'$ is positive on $(-\infty, 0)$ (by Claim 7).

*Case 2*: $d'(0) > 0$. Since $(x^\theta_t, f_t(x^\theta_t)) = (x^\theta_t, y^\theta_t)$ is the unique optimal response to $t$ (by Claim 5), we have $(x^\theta_t, y^\theta_t) \succ_{t,\theta} (0, f_t(0)) = (0, 0)$. So, as the preference $\succeq_{t,\theta}$ is continuity, there exists an $\epsilon > 0$ such that $(x^\theta_t, y^\theta_t + \epsilon) \succ_{t,\theta} (0, 0)$. Hence, as $\succeq_{t,\theta}$ is strictly convex, any strict convex combination of $(x^\theta_t, y^\theta_t + \epsilon)$ and $(0, 0)$ is also strictly preferred to $(0, 0)$. That is, $(\lambda x^\theta_t, \lambda(y^\theta_t + \epsilon)) \succ_{t,\theta} (0, 0)$ for all $\lambda \in (0, 1)$; or equivalently, $(x, x\frac{y^\theta_t + \epsilon}{x^\theta_t}) \succ_{t,\theta} (0, 0)$ for all $x \in (0, x^\theta_t)$. So the function $V : \mathbf{X}_t \to \mathbb{R}$ defined by $V(x) = u_{t,\theta}\left(x, x\frac{y^\theta_t + \epsilon}{x^\theta_t}\right)$ satisfies $V(x) \geq V(0)$ whenever $x < x^\theta_t$. As $V$ is given by $V(x) = v(x - \theta - P(t)) - d\left(x\frac{y^\theta_t + \epsilon}{x^\theta_t}\right)$ and as $v$ and $d$ are differentiable, the chain rule implies that $V$ is differentiable with derivative given by

$$V'(x) = v'(x - \theta - P(t)) - \frac{y^\theta_t + \epsilon}{x^\theta_t} d'(x\frac{y^\theta_t + \epsilon}{x^\theta_t}).$$

As $V(x) \geq V(0)$ whenever $x < x^\theta_t$, we have $0 \leq V'(0)$, i.e.

$$0 \leq v'(-\theta - P(t)) - \frac{y^\theta_t + \epsilon}{x^\theta_t} d'(0) < v'(-\theta - P(t)) - \frac{y^\theta_t}{x^\theta_t} d'(0), \tag{9}$$

where the latter inequality holds by $d'(0) > 0$. In the last expression,

$$\frac{y^\theta_t}{x^\theta_t} = \frac{f_t(x^\theta_t)}{x^\theta_t} \geq \lim_{x \downarrow 0} \frac{f_t(x)}{x} = f'_t(0)$$

by Claims 1 and 2. This inequality and (9) imply that $0 < v'(-\theta - P(t)) - d'(0)f_t'(0)$, i.e. that $0 < U'(0)$, q.e.d.

*Claim 9.* For all $t \in T$, $d'(0)f_t'(0)$ has a unique inverse image $v'^{-1}(d'(0)f_t'(0))$ under $v'$ (denoted $D(t)$).

Let $t \in T$. Uniqueness of the inverse image holds because $v'$ is a strictly decreasing (by Claim 6) and hence one-to-one. To show existence, I have to show that $d'(0)f_t'(0)$ is in the range $v'(\mathbb{R})$. By PM, there exists a type $\theta \in \mathbb{R}$ for whom $(x, y) = (0, 0)$ is not an optimal response to $t$; so the optimum $(x_t^\theta, f_t(x_t^\theta))$ satisfies $x_t^\theta > 0$. Hence, by Claim 8, $v'(-\theta - P(t)) > d'(0)f_t'(0)$. As we also have $d'(0)f_t'(0) \geq 0$ (by Claims 1 and 2), $d'(0)f_t'(0) \in [0, v'(-\theta - P(t))] = [v'(0), v'(-\theta - P(t))]$. So, as $[v'(0), v'(-\theta - P(t))] \subseteq v'(\mathbb{R})$ (because $v'(\mathbb{R})$ is an interval by Claim 7), $d'(0)f_t'(0) \in v'(\mathbb{R})$, q.e.d.

*Claim 10.* For all policies $t \in T$, a type $\theta \in \mathbb{R}$ has $x_t^\theta > 0$ (i.e. is a terrorist) if and only $\theta > -(P(t) + D(t))$.

Let $t \in T$ and $\theta \in \mathbb{R}$. We have $x_t^\theta > 0$ if and only if $v'(-\theta - P(t)) > d'(0)f_t'(0)$. The function $v' : \mathbb{R} \to \mathbb{R}$ is by Claim 6 strictly decreasing, hence is invertible with strictly decreasing inverse function $v'^{-1}$, whose domain $v(\mathbb{R})$ contains both sides of the last inequality by Claim 9. So, applying $v'^{-1}$ on both sides of the inequality yields the equivalent inequality $-\theta - P(t) < v'^{-1}(d'(0)f_t'(0))$ $(= D(t))$, which in turn is equivalent to $\theta > -(P(t) + D(t))$. ■

*Derivations for the example after Theorem 2.* All functions in the example satisfy the required differentiability or convexity conditions. At all $x \leq 0$ we have $v'(x) = (-(-x)^a)' = a(-x)^{a-1}$. So the inverse $v' : \mathbb{R} \to \mathbb{R}$ is at all $z \geq 0$ given by $v'^{-1}(z) = -(z/a)^{1/(a-1)}$. Hence $D(t) = v'^{-1}(d'(0)f_t'(0)) = v'^{-1}(bt) = -(bt/a)^{1/(a-1)}$. To minimise $P(t) + D(t)$, note that at all $t > 0$ this function is differentiable with derivative

$$\begin{aligned}\frac{d}{dt}\left[P(t) + D(t)\right] &= \frac{d}{dt}\left[(t - \bar{t})p - \left(\frac{tb}{a}\right)^{\frac{1}{a-1}}\right] \\ &= p - \frac{(b/a)^{\frac{1}{a-1}}}{a-1} t^{\frac{1}{a-1}-1} = p - \frac{(b/a)^{\frac{1}{a-1}}}{a-1} t^{\frac{2-a}{a-1}},\end{aligned}$$

which is zero if and only if

$$t = \left[\frac{p(a-1)}{(b/a)^{1/(a-1)}}\right]^{\frac{a-1}{2-a}} = \left[\frac{(b/a)^{1/(a-1)}}{p(a-1)}\right]^{\frac{a-1}{a-2}} = \frac{(b/a)^{1/(a-2)}}{(p(a-1))^{(a-1)/(a-2)}}.$$

This only stationary point of $P(t) + D(t)$ is indeed a (global) minimum, since the above expression for $\frac{d}{dt}\left[P(t) + D(t)\right]$ is strictly increasing in $t > 0$ (using that $a > 2$). ■