1 History: deterrence in the Cold War

Deterrence is an old practice in international politics and other areas of behavior. It has been given plenty of thought and study, yet is not easy to understand or explain. The onset of the Cold War provoked enormous interest in deterrence because its role in international politics, particularly at the global level, promised to be critical. However ancient it is in some ways, the greatest part of what we think we know about it was gleaned in the last six decades of systematic thinking and research on deterrence. I won't inflict a lengthy review of its modern history. However, certain comments about deterrence theory and deterrence during the Cold War will be useful for what comes later. I briefly outline what we thought we were doing in managing the Cold War via nuclear deterrence and assess, briefly, the actual role it played in preventing another great war. What the parties *thought* they were doing was not always what they *were* doing and the role of nuclear deterrence was not entirely what it seemed. For those familiar with all this, apologies and a request that you grimace and bear it.

The origins of Cold War deterrence

The essence of deterrence is that one party prevents another from doing something the first party does not want by threatening to harm the other party seriously if it does. This is the use of threats to manipulate behavior so that something unwanted does not occur: "... the prevention from action by fear of the consequences. Deterrence is a state of mind brought about by the existence of a credible threat of unacceptable counteraction" (*Department of Defense Dictionary* 1994). This is fairly straightforward and refers to behavior practiced by nearly all societies and cultures at one level or another.[1] Thus it is hardly surprising to find it used in international politics. However, there the concept came to be applied explicitly, and narrowly, to threats for *preventing an outright military attack*. In a technical sense deterrence is used in international politics in far more ways than this, but forestalling attacks became the focus. Thus a more elaborate definition would be that in a deterrence situation one party is thinking of attacking, the other knows it and is issuing threats of a punitive response, and the first is deciding what to do while keeping these threats in mind (Morgan 1983, pp. 33–42).

Deterrence is distinguished from *compellance*, the use of threats to manipulate the behavior of others so they stop doing something unwanted or do something they were not previously doing.[2] As with deterrence, in security affairs a compellance threat also normally involves military action and often the unwanted behavior to be stopped or steps to be taken involve the use of force, e.g. stop an invasion that has begun, pull out of an occupied area. The distinction between the two is quite abstract; in confrontations they are often present together and virtually indistinguishable. Nevertheless, we attend to the distinction because analysts consider compellance harder than deterrence – it is more difficult to get people/governments to stop doing something they are already doing, like doing, and prepared carefully to do. We now think this is because people tend to be more reluctant, under duress, to take a loss – to give up a benefit in hand – than to forgo seeking an additional benefit of equivalent value. Also, using force to maintain the status quo often seems psychologically more legitimate (to the parties involved and observers) than trying to change it.

[1] Other definitions: "persuasion of one's opponent that the costs and/or risks of a given course of action he might take outweigh its benefits" (George and Smoke 1974, p. 11) – a very broad definition covering almost all forms of influence; "discouraging the enemy from taking military action by posing for him a prospect of cost and risk outweighing his prospective gain" (Snyder 1961, p. 35) – narrows what is to be deterred; ". . . the effective communication of a self-enforced prediction that activity engaged in by another party will bring forth a response such that no gain from said activity will occur, and that a net loss is more probable" (Garfinkle 1995, pp. 28–29) – a very precise, rational-decision conception somewhat at odds with threatening under *nuclear* deterrence to blow the enemy to kingdom come, a real "net loss"; ". . . the absence of war between two countries or alliances. If they are not at war, then it is reasonable to conclude that each is currently being deterred from attacking the other" (Mueller 1989, p. 70). This makes deterrence ubiquitous – everyone is ready to attack everyone else if not restrained – which is not rewarding analytically.

[2] A recent work on compellance, where the distinction is reaffirmed, is Freedman 1998a.

In practice the two overlap. For one thing, they involve the same basic steps: issue a threat, the credibility of which is vital; avoid having the threat make things worse; and thus compel the other side to behave itself. However, parties to a conflict often define "attack" and "status quo" differently, so they disagree over who is attacking whom. When the US threatens military action to halt North Korea's nuclear weapons program is this *deterrence* of a provocative step (a kind of "attack" by the North), or *compellance* of the North to get it to stop what it is doing? And is the US defending the status quo (in which the North has no nuclear weapons) or aggressively threatening the North, which has not directly attacked the US or its allies? The parties in such cases disagree on the answer. If compellance is harder than deterrence then *it matters what the opponent thinks is the situation* since that is crucial to his reaction to the threat. In the example both deterrence (in the US view) and compellance (in North Korea's view) are present.

Thus we should put less emphasis on the distinction between deterrence and compellance and instead treat them as interrelated components of *coercive diplomacy*, the use of force or threat of force by a state (or other actor) to get its own way. This book is about deterrence but assumes that an overlap with compellance is often present and that the two can and must often be discussed interchangeably when examining real-world situations.

In settled domestic societies, deterrence is a limited recourse, used only in particular circumstances and rarely expected to provide, by itself, a viable way to prevent unwanted behavior. In international politics it has had a more pervasive presence. Used primarily as a *tactic*, it has also had a role as a strategic behavior within the jockeying for power that preoccupies states. However, while it was a popular recourse of those fearing attack, it was not the only or even the predominant one, and was not thought of, in itself, as a true strategy.

Without nuclear weapons and the Cold War deterrence would have remained an "occasional stratagem" (Freedman 1996, p. 1). After World War II, for the first time, deterrence evolved into an elaborate *strategy*. It eventually became a distinctive way of pursuing national security and the security of other states or peoples. Nuclear weapons forced those who possessed them, particularly the superpowers, to turn deterrence into a new and comprehensive strategy that touched, shaped, and coordinated many policies and activities. It came to seem intrinsic to international politics, an omnipresent, natural, and continuous recourse in a dangerous environment, something governments engaged in as a regular feature of their existence.

In addition, however, deterrence by the superpowers and their blocs was gradually developed further into *cooperative security management* for the global international system. The superpowers began with unilateral steps to keep safe via deterrence, but the interactions between their deterrence postures soon constituted an elaborate deterrence structure (a "regime"), which constrained them and their allies in numerous ways (not always to their liking) and eventually impelled them into joint efforts to better manage this structure. This had the effect, intended or not, of producing a large increment of security management for the system. Deterrence became a cornerstone of international politics, on which virtually everything else was said to depend.

Thus deterrence came to operate on three levels: as a tactic, as a national security strategy, and as a critical component of security for the international system. Of these the last two made it a suitable subject for theoretical analysis, but it was deterrence as a *national* strategy, in particular within a mutual deterrence relationship, that provided the basis for the theory and became its central focus.

The theory was developed initially to prescribe. The initial question was not "what factors are associated, empirically, with success or failure in deterrence?" but "what are the requirements for a credible deterrence policy?" The straightforward answer (Kaufmann 1954) was to persuade your opponent

(1) that you had an effective military capability;
(2) that it could impose unacceptable costs on him; and
(3) that you would use it if attacked.

The goal was to assist governments to survive in the nuclear age, to conduct an intense conflict without a catastrophe. The stimulus was the appearance and proliferation of nuclear weapons, but in a larger historical context development of some sort of deterrence theory was overdue. Many elements of deterrence thinking appeared before World War II (Questor, 1966; Overy 1992) and important concepts in arms control applied under nuclear deterrence theory were widely discussed after World War I. Nuclear deterrence is best understood as a solution to a fundamental problem of long standing. The evolution of military and other capabilities for war of major states had, well in advance of nuclear weapons, reached the point where great-power warfare, particularly on a systemwide basis involving most or all of the great powers, could be ruinously destructive. One element was the rising destructiveness of weapons. Artillery became extremely accurate at ever longer distances, rifles replaced muskets, machine guns appeared. Other capabilities of military relevance were greatly enhanced. Vast increases in productivity, combined with new bureaucratic and other resources, gave great states huge additional capacities for war. They acquired greater abilities to sustain and coherently manage large forces and exploited the breakthroughs in communications and transportation. Nationalism added the collective energies of millions, whether for raising armies and money or for production of everything those forces needed (Levy 1982, 1989b). Great states became capable of huge wars – in size of forces, levels of killing and destruction, duration, and distance. This was foreshadowed by the Napoleonic Wars, displayed by the American Civil War, and grasped in Ivan Bloch's penetrating analysis at the turn of the century of what the next great war would look like (Bloch [1899] 1998). All that was missing was a graphic example, which World War I supplied. It had become possible to conduct "total war."[3]

It is important to understand just why this was *the* problem. For practitioners of international politics it was not war itself. Particularly for great powers, war had always been a central feature of the international system, a frequently used and legitimate tool of statecraft, the last recourse for settling disputes, the ultimate basis for the power balancing that sustained the system and the members' sovereign independence. It had also been fundamental in creating nation states. "From the very beginning the principle of nationalism was almost indissolubly linked, both in theory and practice, with the idea of war," and thus "It is hard to think of any nation-state, with the possible exception of Norway, that came into existence before the middle of the twentieth century which was not created, and had its boundaries defined, by wars, by internal violence, or by a combination of the two" (Howard 1991, p. 39). It was difficult to imagine international politics without war since it seemed an inevitable adjunct of sovereign autonomy. War had last threatened to get completely out of hand during the Thirty Years War (1618–48) and states

[3] Vastly destructive wars are not unique to the twentieth century (Ray 1989; Mueller 1989, pp. 3–13). But beginning in the nineteenth century the capacities for destruction, even in a losing effort, expanded rapidly with the developments listed above and others (such as conscription) which ". . . served to make it much more likely that war, when it did come, would be total . . . The closely packed battle, in which mass is multiplied by velocity, became the central feature of modern European military thought. For the first time in history, governments were coming into possession of constantly expanding means of waging absolute war for unlimited objectives" (Dougherty and Pfaltzgraff 1981, p. 195).

had responded by setting the Westphalian system into operation partly to get it under control. In the twentieth century the system was again being overwhelmed by war. Detaching sovereign rule, which is highly prized, from the rampant use of force for selfish purposes is the ultimate security problem of international politics, and now it threatened to destroy everything.

The development of deterrence was driven by a particularly onerous alternative solution that had emerged some time earlier to the threat of great-power war. Confronting the distinct possibility that the next war would be enormously destructive and costly, states worked hard to devise variants of a *cheap-victory strategy*. The idea was to ensure that the great costs, destruction, and loss of life would fall mainly on the other side. This was foreshadowed in Napoleon's shattering victories via single grand battles that collapsed the opponent. It dominated Prussia's wars against Denmark, Austria, and France in 1862–1871, wars so successful that such strategies have shaped military planning ever since. The Prussian approach involved diplomatically isolating the opponent, then utilizing industrial-age resources and nationalism to mobilize rapidly and throw huge well-coordinated forces into the initial battles to overwhelm the opponent, inflicting a complete defeat to end resistance. As a result, the major states approaching World War I had plans for rapid mobilization and decisive offensive thrusts to overwhelm the opponent in the opening battles, forcing the enemy to collapse in just weeks before intolerable casualties and costs were incurred. The Schlieffen Plan sought a cheap victory, as did the French Plan 17, the prewar plans of Russia and the Austro-Hungarian Empire, and British plans for fighting with Germany on land. Hitler sought to recapitulate the Prussian approach by isolating the target state and inflicting a (blitzkrieg-style) defeat so as to avoid a long and costly war. The point of the French Maginot Line was to fight a cheap, minimal-casualty war by exploiting the superiority of settled defenses (supposedly demonstrated in World War I) to wear down the attacking Germans; eventually France would push into a gravely weakened Germany and impose defeat at little cost. Japan's attacks at Pearl Harbor and elsewhere in late 1941 were meant to establish an impregnable defense far from home that would wear out the Americans and bring a settlement on terms favorable to Japan, producing victory at low cost.

Cheap-victory solutions influenced the development of deterrence theory in two broad ways. In the twentieth century these strategies were terrible failures in the world wars. Often initially successful, in the end they failed and the resulting wars were dreadful even for the winners, making the problem of great-power warfare clear to everyone. Another approach was obviously needed. In addition, cheap-victory solutions – which often turned on winning quickly – could be highly destabilizing because they usually required striking by surprise or before the other party was fully prepared. Thus once a war looked quite possible they could have the effect of initiating it.

Making interstate war virtually impossible by either fundamentally changing international politics or abandoning it was difficult to contemplate. Neither seemed remotely feasible so serious thinking shifted, almost inevitably, toward how great wars might be deterred, temporarily or indefinitely. The first effort along these lines was the formation of the League of Nations. It was meant to provide *collective actor deterrence* – deterrence by the entire membership against any member thinking about attacking another. In addition, components of what would become deterrence theory began to emerge in the 1920s. Analysts began to describe certain forces and capabilities as dangerous because they made war by surprise attack or on short notice plausible. Hence the ban on conscription imposed by the winners on the losers after World War I; the absence of masses of trained men, plus limits on the size of the losers' professional forces, would – it was hoped – prevent the quick mobilization of vast armies to achieve a cheap victory. Analysts began to characterize offensive, as opposed to defensive, forces and postures as too provocative. The British eventually developed strategic bombing as a deterrent, with preliminary thoughts on how key targets might be industrial and military or the will, politically and psychologically, of the enemy to continue to fight, foreshadowing the distinction between deterrence via defense (war-fighting) and deterrence via punishment (retaliation). US military thinking was similarly interested in deterrence through the threat of strategic bombing (Overy 1992).

What drove these efforts to coalesce into a theory was the coming of nuclear (particularly thermonuclear) weapons and the emergence of more than one national nuclear capability, especially when linked to ballistic missiles. Those weapons seemed ideal for achieving a cheap victory and thus were regarded (by thoughtful scientists even during their development) as very destabilizing. And they promised destruction at even higher levels.

Nuclear deterrence was the ultimate in threatening awful consequences to prevent wars. It had been known for years that great-power wars could be awful, so threatening one was not, in itself, new. The innovation lay in using nuclear weapons to *strip any cheap-victory strategy of plausible success*, to leave an opponent no reliable way to design a great-power war in which it would suffer little and gain much. As this is important for the discussion later on it is worth emphasizing. It was not that nuclear weapons promised so much destruction that made them crucial in deterrence, it was that they made this destruction seem virtually *unavoidable under any plausible strategy*. This was the crux of the "nuclear revolution" (Jervis, 1989a) in statecraft.

The essence of deterrence theory

In discussing the *theory* it is important to distinguish it from deterrence *strategy*.[4] Deterrence strategy refers to the specific military posture, threats, and ways of communicating them that a state adopts to deter, while the theory concerns the underlying principles on which any strategy is to rest. Failure to keep this in mind is largely responsible for the frequent but mistaken suggestion that there are many theories of deterrence. Mostly there are different strategies, not theories. The strategies vary in how they operationalize key concepts and precepts of the theory. As for alternative theories, they are mostly theoretical fragments, not theories.[5]

The key elements of the theory are well known: the assumption of a very severe conflict, the assumption of rationality, the concept of a retaliatory threat, the concept of unacceptable damage, the notion of credibility, and the notion of deterrence stability. Examined briefly here, each is of importance later in considering whether deterrence has changed since the Cold War.

Severe conflict

Since deterrence theory was developed to help states cope with the Cold War, the nature of that struggle had great impact on it. The most

[4] Standard works on deterrence theory are: Freedman 1981; George and Smoke 1974; Jervis 1979, 1984, 1989b; Morgan 1983; Powell 1990; Questor 1986; Maxwell 1968; Wohlstetter 1959; Brodie 1959; Kahn 1961, 1965; Schelling 1960, 1966; Snyder 1961; Mearsheimer 1983; Jervis, Lebow and Stein 1985; Lebow and Stein, 1989, 1990a, 1994; Lebow 1981; Stein 1991.

[5] Escalation dominance/war-fighting was not a different theory, as is sometimes suggested. During the Cold War it presented a different view of what generates stability and credibility, what was unacceptable damage (particularly for Soviet leaders), and how to cope with deterrence failure. Colin Gray (1990, p. 16) says "theories of deterrence – or approaches to theories – are the product of their time, place, and culture," but the operationalization shifts more than the theory itself.

important feature in this regard was its *intensity*. To both sides it was total and ultimate, with the future of the world at stake. Both considered war a constant possibility; the enemy would not hesitate to attack if a clear chance for success arose. Thus deterrence had to be in place and working all the time, every day. The necessary forces had to be primed and ready to go. All that was keeping this conflict from turning into a war, probably total war, was deterrence. It stood between the great states and Armageddon.[6]

Years ago I devised the distinction between "general" and "immediate" deterrence. In general deterrence an actor maintains a broad military capability and issues broad threats of a punitive response to an attack to keep anyone from seriously thinking about attacking. In immediate deterrence the actor has a military capability and issues threats to a specific opponent when the opponent is already contemplating and preparing an attack. Thus an immediate deterrence situation is a crisis, or close to it, with war distinctly possible, while general deterrence is far less intense and anxious because the attack to be forestalled is still hypothetical. For years the Cold War was conducted as if we were on the edge of sliding into immediate deterrence. The attack-warning systems operated continuously, weapons and forces were on high alert, and there were elaborate calculations as to whether the opponent could pull off a successful preemptive attack or had programs under way to produce such a capability. It seemed that a crisis could erupt quite suddenly and lead to war, and there were very strong threat perceptions. One Strategic Air Command (SAC) commander testifying in 1960 on why his bombers should be constantly on airborne alert said: "...we must get on with this airborne alert to carry us over this period. We must impress Mr. Khrushchev that we have it and that he cannot strike this country with impunity. I think the minute he thinks he can strike this country with impunity, we will 'get it' in the next 60 seconds" (Sagan 1993, p. 167).

This was a distorted and distorting perspective. Seeing the opponent as just looking to attack, as "opportunity driven," was a Cold War political assessment of a particular challenger. There is no necessity to start with this assumption – we did so because that is where, at the

[6] In the Soviet bloc the stakes seemed just as high, the enemy just as ruthless and willing to use war, but war seemed much less likely to come at any moment. Soviet strategic forces were less often on alert; political portents of a Western attack were expected to provide enough warning in advance. Only under Yuri Andropov, confronting the Reagan Administration, did Moscow consider an attack at almost any moment a real possibility.

time, deterrence *strategy* had to start. The theory worked outward from considering how to cope with a war-threatening confrontation, a worst-case analysis, rather than with general deterrence and working down to the rare and extreme situation of an impending war. Immediate deterrence was the primary consideration, dominating most thinking even about general deterrence. This was awkward because much of the theory, therefore, particularly in connection with arms control, came to be concerned with stability in situations in which neither party wanted a war. Refinements emphasized, in spiral-model fashion, the existence of a conflict and the nature of military plans and deployments as potential causes of war in themselves, and not only the machinations of an opportunistically aggressive opponent.[7]

This had a strong effect on theory and strategy. It is hard to imagine the theory as we know it ever having emerged if each side in the East–West dispute had felt the other had little interest in attacking. The theory could operate as if deterrence was critical for preventing an attack. It did not explore what the motivations for war might be (and thus whether they were always present). It simply took as its point of departure a conflict so intense that the two sides would likely go to war if they thought they could get away with it. Hence the recurring concern in the US that deterrence was delicate and could easily be disturbed by developments that might seem to give the other side a military advantage. Deterrence strategy, as Lebow and Stein (1990a) emphasize, came to view the occurrence of war as related to windows of opportunity generated by a flawed deterrence posture. More precisely, theory and strategy operated on the expectation that each side must assume the other would attack if a suitable opportunity emerged. (Actually, the theory did not have to do this – it simply concerned what to do to deter if and when a state faced a possible attack.)

This was why the theory paid little attention to other ways of preventing war, such as by seeking to reconcile differences or offering reassurances and incentives. Efforts to suggest how deterrence might be used in conjunction with other approaches to peace were never incorporated; instead, it was about preventing a war when these other approaches had failed or could not be expected to work. In one sense this was fine. The theory was not held to be comprehensive or depicted as the only route to security under any circumstance. It merely

[7] Thus Jervis (1976) contrasts a "deterrence model" with a "spiral model" when the theory embraced both.