Chapter 1

Introduction—The Evolution of Deterrence Strategy and Research

Lawrence Freedman

Contents

1.1 Introduction.....	1
1.2 The Cold War Focus: General Nuclear Deterrence.....	3
1.3 Moving on: Including Conventional Deterrence	5
1.4 Beyond the Cold War and General Deterrence	6
1.5 The Enduring Relevance of Deterrence Strategy and Research	8
References	9

1.1 Introduction

The concept of deterrence has dominated Western strategic thought for some seven decades. It shows no signs of easing its grip. In the face of any new security threat, such as terrorism or cyber-attacks, one of the first questions to be asked is ‘can this be deterred?’ Even when the answer is not very encouraging the inclination is to persevere until some way is found at least to reduce if not remove the threat through some form of deterrence. This may have less to do with deterrence’s reliability or effectiveness as a strategy and more because of its inherent normative appeal. When a state adopts a deterrence strategy it signals that it does not seek a fight but still considers some interests to be so vital that they are worth fighting for. It implies a defensive intent without weakness. It seeks to prevent aggression while being non-aggressive. It sustains rather than disrupts the status quo. For these reasons, it has positive associations that other potential strategies lack. Appeasement as a

L. Freedman (✉)

Department of War Studies, King’s College London, London, UK

e-mail: lawrence.freedman@kcl.ac.uk

© The Author(s) 2021

F. Osinga and T. Sweijts (eds.), *NL ARMS Netherlands Annual Review of Military Studies* 2020, NL ARMS, https://doi.org/10.1007/978-94-6265-419-8_1

deliberate strategy has been discredited since the 1930s; conquering other states is now seen to be as demanding as it is illegal. There is no dishonour in deterrence.

The main objections to deterrence strategies are because they provide the core rationale for possessing nuclear weapons. Those arguing for nuclear abolition often argue that the deterrence effect is a chimera. What were thought to be deterrent successes either had other causes or could be achieved by other means. This can lead to playing games with history to make a point. It was of course entirely possible that there would have been no Third World War even if nuclear weapons had never been invented but in the post-1945 world at crucial points nuclear weapons acted as a vital source of restraint.¹ A stronger, more moderate argument is that nuclear deterrence was a thin reed upon which to rely and might have let governments down at crucial moments. But that was an argument about the limits of deterrence and not its potential validity. There is nothing effortless about deterrence. It demands close attention to how threats are designed, conveyed and, if necessary, implemented.

The concept itself is simple enough. Deterrence occurs when A persuades B not to take a specified step by convincing B that whatever the anticipated gains the likely costs will be higher. When A issues a threat, its effectiveness will depend on B's perception of what it might mean as much as A's intent. If B does not take A's threat seriously and concludes that it can be safely ignored then deterrence will fail. Or A may fail to deter through negligence. It knows that B needs deterring but does not realise exactly what B is up to until too late and so is caught by surprise. Once A needs to retrieve a lost position deterrence has become irrelevant. The tables may be turned as B is now deterring A to preserve a new status quo. Nor is there a standard formula suitable for application in any situation where deterrence is required. What might work when vital interests are involved might not work when the stakes are low. A's stern threats might hit home when B is paying attention but miss completely when B is distracted or if there is a lot of background noise. What worked last time might not work this time, not least because B knows what to expect. And, as deterrence depends on the status quo holding, when nothing much changes can we be sure that this is because of the deterrent threat? Is that why B has held back? Perhaps no hostile action was ever intended. Or if it has not happened that might be for reasons unrelated to deterrence. Deterrence is therefore simple in principle and a natural strategy to adopt but it is not so straightforward when it comes to implementation. It poses both a conceptual and a practical challenge.²

¹Mueller 1989; Gaddis et al. 1999.

1.2 The Cold War Focus: General Nuclear Deterrence

The Cold War flattered deterrence. It was credited with the success of the West in containing the Soviet threat. It was also during the Cold war that deterrence acquired a compelling conceptual framework. The idea of deterrence had a long history. Its origins lay in criminology. The utilitarian philosopher Jeremy Bentham supposed that criminals were sufficiently rational and self-interested to calculate when the costs of punishment would outweigh the potential benefits of crime. On this basis he proposed the term ‘determent’.² The same idea was understood in international affairs even when the word was not used. The antecedents of Cold War thinking can be found in debates about how to deal with the prospect of mass air raids during the 1930s.³ But it was nuclear weapons that made the difference, especially once the Soviet Union tested an atomic device in August 1949 and as both superpowers moved to ‘city-busting’ thermonuclear weapons. After this point, even when military planners and civilian think-tankers tried to think of clever ways of employing nuclear weapons to win wars, prudence kept on pulling policy-makers back to deterrence.

During the 1950s and into the 1960s most of the important conceptual work on deterrence was undertaken. At the start of the thermonuclear age there was a natural assumption that two equivalent, deadly multi-megaton arsenals would lead to a ‘balance of terror’ which would lead to an uncomfortable but potentially durable peace. But what if the ‘balance of terror’ was not so stable? In the mid-1950s analysts at the RAND Corporation demonstrated that a well-designed first-strike directed against the opponent’s nuclear weapons might deny it the chance to retaliate. To guard against being caught out in this way a second strike capability was required—the ability to absorb a first strike and still retaliate.⁴ Such thinking encouraged a technological arms race, for defensive as much as offensive reasons. Vulnerability to a first strike might lead to vulnerability to political pressure. Eventually second-strike capabilities won out over first-strike capabilities, largely as the result of the development of relatively invulnerable ballistic missile carrying submarines.

Another early response to the problem of the balance of terror, prompted by the Korean War, was to accept that this might preclude total war but then suggest that limited wars might still be possible.⁵ Out of the exploration of these possibilities came the notion of escalation. This was at first a tragic concept, suggesting that any serious fighting between the superpowers, even if at first limited, would soon erupt into total war. This concept came about as a critique of the proposition that even if low-yield, short-range ‘tactical’ nuclear weapons were used, a war could still be kept ‘limited’. In the 1960s escalation was presented in more positive terms. Instead

²Bentham 1830.

³Quester 1966.

⁴Wohlstetter 1959.

⁵Osgood 1957.

of a quick route to tragedy the idea of an 'escalation ladder' was used to show how a war might expand in stages.⁶ At least in the early stages it would be possible to control the process. The basic idea was to establish sufficient dominance at one step on the ladder to put the onus on the adversary to take the risk of moving to the next step, with the violence becoming more destructive and less controllable. This was therefore intra-war deterrence. The mutual danger had been insufficient to stop the onset of war but still sufficient to encourage caution when the escalatory process was pointing to total war. This turned war into a competition in risk-taking, with deterrence failing by degrees. At some point escalation dominance would become impossible. There would no longer be a way to limit nuclear use or discriminate in targeting. The only remaining possibility was utter catastrophe.

The fear of escalation and uncertainty over the stability of the balance combined to generate anxieties that war might come about, and therefore deterrence might fail, not out of deliberate choice but because of miscalculation or even system malfunctions. A rogue order to attack or a faulty early warning system could set in motion terrible events. Such fears encouraged the idea that the superpowers should find ways to cooperate in order to reduce the risk of inadvertent escalation.⁷ Arms control was a way of agreeing on how to structure nuclear forces in order to bring stability to the balance. In this way the notion of Mutually Assured Destruction was embraced and consolidated in arms control agreements, describing a situation in which nuclear exchanges would be unavoidably catastrophic for all belligerents. The superpowers had to accept that they had no route to victory in a nuclear war.

This left an awkward question, especially for the United States and its NATO allies. The original idea behind the alliance was that the US would not wait, as it had done in the previous two world wars, to come to the aid of the western democracies if they were attacked. As the Soviet Union and its satellite states enjoyed conventional superiority deterrence was assumed to depend from the start on America's readiness to initiate nuclear war. The need therefore was to deter all war and not just nuclear war. But if mutual assured destruction meant that the nuclear arsenals neutralised each other, then might the Soviets not feel free to see what could be done with conventional war without risking the devastation of their homeland? How could US nuclear threats be credible when any implementation risked retaliation in kind? How could allies be confident in the US nuclear umbrella if that meant an American government must put American cities at risk to protect European cities? This was the problem of extended deterrence.⁸ It was one thing if nuclear arsenals were geared solely to national defence and nothing else, which was the original French concept. That had a sort of credibility. It was quite another to prepare to wage nuclear war on behalf of third parties. The obvious credibility problems surrounding US nuclear guarantees were the greatest stimulus to

⁶Kahn 1965.

⁷Schelling and Halperin 1961.

⁸Bobbitt 1988.

creativity in deterrence theory. It led to a number of distinctive and sometimes contradictory lines of thought that still influence thinking about deterrence.

The most influential of these lines of thought distinguished between deterrence by denial and deterrence by punishment.⁹ This followed from the basic definition of deterrence as persuading an adversary that prospective costs would outweigh prospective gains. Because of the nuclear association deterrence was presumed to work through the threat of severe punishment. But this definition allowed for a completely different approach based on denying the enemy gains. In the NATO context deterrence by denial came to be associated with conventional capabilities.¹⁰ If the Warsaw Pact could not mount an effective invasion then NATO would need to rely on the threat of nuclear first use. Against this it was argued that conventional deterrence would be expensive while the lack of a nuclear dimension might encourage exploratory Soviet aggression. With the end of the Cold War the problem of dependence on nuclear first use switched from NATO to Russia that now had to cope with conventional inferiority. While the alliance could not quite bring itself to move to a no first use promise the assumption in practice was that conventional superiority would deter most forms of aggression against NATO countries.

1.3 Moving on: Including Conventional Deterrence

Conventional deterrence was much more credible in principle and also had many more potential applications. There were still a few—and generally extreme—contingencies when it would be appropriate to talk about nuclear threats but there were no similar restrictions with conventional capabilities. The shift to conventional denial opened up a whole range of deterrence possibilities that were of no interest before. Nor was it necessary to confine conventional forces to denial—as weapons became more accurate over long ranges they could be used to inflict tailored punishments as well. Liberating deterrence from its nuclear associations also made the concept analytically more interesting. Explicit nuclear threats were few and far between but in principle there were numerous instances of conventional deterrence, going well back into history. This made possible what Robert Jervis called the ‘third wave’ of deterrence theory based on empirical case studies.¹¹ These tended to be instances where it could be shown that one side was prepared to act, and another had tried to deter, sometimes with success and sometimes without. By comparing many cases it might be possible in principle to see what factors made deterrence more or less likely to succeed.

⁹Snyder 1961.

¹⁰Mearsheimer 1983.

¹¹Jervis 1979, pp. 289–324.

Whether or not these helped illuminate the problems of Cold War deterrence was another matter. In this context, whatever reliance might be placed on conventional denial the possibility of nuclear punishment was always present. Whatever the doubts about the credibility of extended nuclear deterrence it was hard to ignore the possibility that with so many nuclear weapons deployed, often as part of army, air and naval formations, there was always a risk that in the heat of battle some might be launched. This moved the appreciation of the deterrent effect of nuclear weapons away from the credibility and specificity of threatened use and toward the residual risk that even if policy-makers were desperate for this not to happen, they might nonetheless be used. It might be hard to describe a chain of events that would lead to a rational decision to initiate nuclear war but it was impossible to preclude its possibility once a major war had begun. Despite the efforts to imagine controlled steps up an escalation ladder it was much easier to imagine how the more tragic concept of escalation would influence events. It continued to remind of how the best efforts to keep a war limited might be dashed as matters got out of hand. Nuclear use would be propelled to the fore by the passions and uncertainties of a bitter conflict, leading to a terrible conclusion. This prospect on its own should lead to caution and restraint at a time of crisis. All that was necessary for a deterrent effect was for nuclear weapons to exist in a usable form. This was described as 'existential deterrence'.¹² It was not one side's threats of action that deterred but the risk of an event in which such escalation might occur.

In the circumstances of the Cold War this worked. One reason for this was that there were no obvious flashpoints, at least in Europe, once the anomaly of Berlin was sorted out, first in August 1961 by the construction of the wall that divided the city and then at the end of that decade by a set of agreements encouraged by West Germany's 'Ostpolitik' leading to a détente in Europe. Instead of a direct attack by one alliance against the other scenarios for future war tended to postulate unrest within the satellite states of the Warsaw Pact or else a crisis imported from the Middle East where matters were more dynamic and fluid. Alternatively, anything that might threaten the cohesion of NATO, including an American decision to withdraw its forces from Europe on the grounds that they were no longer needed, potentially risked unsettling the status quo.

1.4 Beyond the Cold War and General Deterrence

From this three large conclusions might be drawn about deterrence. The first was that it was not good enough to consider it in terms of a particular configuration of forces and articulation of threats. Deterrence relationships had to be given context. The state of affairs at risk from aggression or disruption had to be addressed along with the national interests of the key actors. In situations marked by turbulence and

¹²McGeorge 1983.

volatility, and with so much going on, identifying the specific move that needed to be deterred might not be straightforward. The second was to recognise the importance of alliance as a source of deterrence. When NATO was formed, the deterrent effect was the result of the United States becoming committed to the security of the West European democracies. So long as that remained the case aggression against Western Europe would be high risk. Should the commitment be withdrawn the situation would change dramatically. The Europeans would need to look to other means to resist Soviet pressure. The third conclusion was that because the European situation stabilised, with neither NATO nor the Warsaw Pact in disarray and no crisis forcing matters to a head, other than occasional flurries of anxiety, deterrence became embedded in the thinking of all key actors and lost all sense of urgency. It became internalised. The preparations for war continued as part of the routines of alliance but political leaders saw no need to put their countries on a war footing.

Patrick Morgan made an important distinction between general and immediate deterrence.¹³ With general deterrence the situation is one in which relations between states are still antagonistic yet the antagonism has long lost its edge. With immediate deterrence the antagonism is sharp and dangerous and A must act at once to deter B's likely aggression. Most deterrence relationships start with an immediate crisis. If it is managed successfully the threats and forces may still stay in place until eventually a point is reached when a resumption of crisis conditions appears unlikely. This should allow underlying political relations to improve. If on the other hand the conflict shows signs of resuming it should be possible to provide timely reminders of why aggression is still a bad idea. Much of the empirical literature is dominated by instances of immediate deterrence. By definition these are situations when not only has deterrence not been internalised but it is barely working at all, for that is why the crisis has occurred. General deterrence, when there is an embedded expectation that nothing much will happen, described a truly successful deterrence strategy. The would-be aggressor not only holds back for the moment but also then stops thinking of aggression as a serious option.

Focussing on the political context helps explain what happened to deterrence after the end of the Cold War. The East-West conflict had reached its own equilibrium between two superpower-led alliances. Both appreciated the possibility of mutually assured destruction and as a result had internalised deterrence. This equilibrium was lost as one alliance collapsed (with most of its members then joining NATO) leaving Russia feeling more vulnerable and insecure. Political relations generally became more complex and fluid, so it was less clear where deterrence was needed and how it might be achieved, especially as non-state actors, including vicious terrorist groups grew in importance. Deterrence was no longer required to solve one big problem. Instead it was called in aid to solve many small ones. Instead of the dominant strategy it was a stratagem available for addressing a

¹³Morgan 1977.

variety of contingencies, some quite unique, using a variety of means, non-military as well as military, although without confidence that it could solve any. This was described as ‘fourth-generation deterrence’, much broader than the earlier generations reflecting the changes in the international system, and so lacking the theoretical coherence or consistent sense of purpose.¹⁴

1.5 The Enduring Relevance of Deterrence Strategy and Research

The big problem of a great power war of course did not go away. Indeed, after a relatively relaxed period it returned, with Russia using force to look after its interests, first in Ukraine and then in Syria, and China flexing its muscles in the Asia-Pacific region. This led to a degree of continuity. The same questions were raised as they had been during the Cold War about the durability of America’s alliances and whether nuclear weapons have a dampening effect on tendencies towards open warfare. There were now other actors whose nuclear arsenals needed to be taken into account, including India, Pakistan, Israel and North Korea. In this way the fourth-generation of deterrence was shaped by the ‘second nuclear age’.¹⁵ A secure second-strike capability was still seen as vital to the practice of nuclear deterrence. There was now an added concern, which had been growing steadily since the 1970s, about the vulnerability of command and control systems. It may not matter if all sorts of enemy military capabilities, including nuclear weapons, are left if a first strike directed against the ‘national command authority’ has left the enemy brainless and paralysed, although it would need considerable confidence to be sure that a decapitation attack would leave the enemy so brainless and paralysed that it would be completely unable to take retaliatory action. Such attacks also raised the question of how a peace can be arranged if there was no one left with whom to negotiate. Yet it is an issue that bothers military planners, especially when governments might be taken out by non-nuclear systems, such as hypersonic weapons. It is also an issue that is not confined to major war. The targeted killing of leaders has become a feature of counter-terrorism and counter-insurgency campaigns. The aim has been more to reduce the effectiveness although it might give someone an opportunity to take on a leadership role pause for thought!

As major war is still best avoided attention has moved to the so-called ‘grey area’ between a comfortable peace and serious fighting, involving proxies, information campaigns and cyber-operations. In the grey area it may be hard to attribute actions to particular actors. So while denial might work in that a degree of resilience and protection can be built in to any system that might be attacked, punishment is

¹⁴Knopf 2010.

¹⁵Bracken 2000.

less straightforward unless the guilty party can be identified with accuracy and some appropriate sanction identified.

Deterrence works best with unambiguous red lines, established over time, linked with vital interests, and backed by clear and credible messages, reinforced by known capabilities, about what will happen if they are crossed. It will work less well as more uncertainties are introduced—about where the lines actually are, how much any transgressions will actually matter, whether there will be much of a response if they are crossed and what difference they will actually make. A decades-long stand-off in the centre of Europe between two great alliances was one thing: sudden crises emerging out of a complex, multi-faceted and fast changing set of political relationships is another. If only for its presentational advantages, deterrence will continue to be seen as the ideal response to most types of security threats. In some situations it should work well—often so well that it is taken for granted. But it would be unwise to play down the challenges of making deterrence work when threats have to be constructed in a hurry to deal with one-off situations with lots of unique complications, amid expressions of doubt and dissent about whether they could or should be acted upon.

References

- Bentham J (1830) The Rationale of Punishment (derived from manuscripts by Bentham written in the 1770s). In: Heward R (ed), London. <http://www.la.utexas.edu/labyrinth/rp/index.html>
- Bobbitt P (1988) Democracy and Deterrence: The History and Future of Nuclear Strategy. Macmillan, London
- Bracken P (2000) The Second Nuclear Age. Foreign Affairs (January/February 2000)
- Gaddis J L, Gordon P H, May E R, Rosenberg J (1999) Cold War Statesmen Confront the Bomb: Nuclear Diplomacy Since 1945. Oxford University Press, Oxford
- Jervis R (1979) Deterrence Theory Revisited. World Politics XXXI.2:289-324 (January 1979)
- Kahn H (1965) On Escalation: Metaphors and Scenarios. Pall Mall, London
- Knopf J W (2010) The Fourth Wave in Deterrence Research. Contemporary Security Policy, 31.1
- McGeorge B (1983) The Bishops and the Bomb. The New York Review of Books (16 June 1983)
- Mearsheimer J (1983) Conventional Deterrence. Cornell University Press, Ithaca
- Morgan P (1977) Deterrence: A Conceptual Analysis. Sage Publications
- Mueller J (1989) Retreat from Doomsday
- Osgood R E (1957) Limited War: The Challenge to American Strategy. The University of Chicago Press
- Quester G (1966) Deterrence Before Hiroshima: The Influence of Airpower on Modern Strategy. Wiley, New York
- Schelling T, Halperin M (1961) Strategy and Arms Control. Twentieth Century Fund, New York
- Snyder G (1961) Deterrence and Defense. Princeton University Press, Princeton NJ
- Wohlstetter A (1959) The delicate balance of terror. Foreign Affairs XXXVII.2 (January 1959)

Lawrence Freedman (CBE) has been Professor of War Studies at King's College London since 1982, and Vice-Principal since 2003. He was educated at the Universities of Manchester, York and Oxford. Before joining King's, he held research appointments at Nuffield College Oxford, IISS and the Royal Institute of International Affairs. Elected a Fellow of the British Academy in 1995 and awarded the CBE (Commander of the British Empire) in 1996, he was appointed Official Historian of the Falklands Campaign in 1997. He was awarded the KCMG (Knight Commander of St Michael and St George) in 2003 and in 2009 was appointed to serve as a member of the official inquiry into Britain and the 2003 Iraq War.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.