Deterrence Theory, Then and Now: There is No Going Back

Frank C. Zagare

ABSTRACT

Perfect Deterrence Theory is a general theory of interstate conflict that is applicable to a wide variety of real-world circumstances, including acute crises and both conventional and nuclear deterrence relationships. The theory is comprised of a set of closely related game-theoretic models that are explored under conditions of complete and incomplete information. Like any theory, Perfect Deterrence Theory's major propositions and policy recommendations are contingent on its defined terms and concepts. The failure to take note of these definitions when evaluating the theory not only does a disservice to the theory, but it also undermines the academic enterprise itself.

During the 1950s and early 1960s, theoretical and policy debates among security study specialists focused squarely on issues that involved deterrence. To maintain at least the semblance of objectivity, many of the most prominent deterrence theorists used game theory to confront these questions in the abstract. Policy analysts, however, made little attempt to hide the fact that their central concern was to formulate defence postures that would enable the United States to deter a Soviet attack on itself and its NATO allies.

The body of literature that emerged from these debates is sometimes referred to as "rational" or "classical" deterrence theory. The labels suggest convergence around a common set of concepts, assumptions, theoretical constructs, and policy recommendations. Of course, there were dissenters, as there always are, and unresolved arguments, as there always will be. Nonetheless, as time passed, classical deterrence theory emerged as a recognizable subfield unified by both a commonality of focus and generally compatible theoretical and policy propositions.

I have argued elsewhere, and at great length, that the classical formulation suffers from both logical inconsistency and empirical inaccuracy,[1] two shortcomings that are fatal to the health and well-being

Frank C. Zagare, "Deterrence Theory, Then and Now: There is No Going Back," *St Antony's International Review* 9, no. 1 (2013): 157-167.

of any theory. I am also largely responsible for developing an alternative theoretical specification called Perfect Deterrence Theory that I claim, rightly or wrongly, corrects for the deficiencies of classical deterrence theory.[2] Using this theoretical framework as a guide, Stephen Quackenbush explored the strategic implications of national missile defence systems.[3] And, in a subsequent review of current theory, he observed that his "application of perfect deterrence theory demonstrates that national missile defence can enhance deterrence stability."[4]

It is largely on the basis of Quackenbush's comment about the efficacy of national missile defence systems that Fred H. Lawson situates Perfect Deterrence Theory in the context of a debate in the earlier literature between proponents of "classical deterrence" and advocates of "defence."[5] Perfect Deterrence Theory, Lawson claims, falls on the wrong side of the argument along with theorists like Albert Wohlstetter, Henry Kissinger, and Herman Kahn who argue that atomic and, later, nuclear weapons were usable in war and that, therefore, defensive systems and deployments were called for.

It is an interesting argument, especially since Stephen Walt has previously suggested that Perfect Deterrence Theory is simply a reinvention of "the central elements of [classical] deterrence theory."[6] It is also wrong. Perfect deterrence theory neither reinvents classical deterrence theory, as Walt charges, nor advocates using nuclear weapons in war and/or uncategorically prescribes national defence systems as an effective mechanism for stabilizing the status quo. Rather, it explores the theoretical implications of a variety of strategic environments, including those where the credibility of deterrent threats are less than perfect (i.e. are uncertain), where war costs may range from low to high (i.e. where the capability or hurtfulness of deterrent threats are not constant) and where the dissatisfaction of states with the existing order is not taken as a given (i.e. where all states are not undifferentiated).[7] It is therefore a general theory of interstate conflict that is easily calibrated to take account of a wide variety of hypothetical or empirical circumstances, only one of which Quackenbush explores in his discussion of national missile defence systems. Such subtleties seem to be beyond the reach of the author of "Back to the Future." I can only speculate as to why.

One reason may be that Professor Lawson has not taken the time to fully understand the theory he so cavalierly attacks. Evidence for this explanation lies in the fact that I have elsewhere staked out a somewhat different position than Quackenbush on the efficacy of national missile defence systems.[8] Specifically, I suggested that there were conditions under which such a system would not seriously undermine strategic stability,[9] but that it would likely do so in the long run if its deployment (by the

United States) antagonized a rival state (either Russia or China).[10] Thus, it is incorrect to suggest or conclude that Perfect Deterrence Theory is deficient simply because it provides a nuanced understanding to a policy question that runs counter to the prescriptions of some classical deterrence theorists and, presumably, of Professor Lawson.

It is also important to point out that Quackenbush notes that his conclusions stem from an *application* of the theory, not from the theory itself.[11] This is not a trivial point. Perfect Deterrence Theory is a general theory that can be used to explore the strategic implications of a wide variety of theoretical or empirical conditions. Quackenbush's application begins with an argument about what those conditions are and draws conclusions about their consequences using the theory as a guide. The theory itself is agnostic about what conditions govern any strategic relationship. Thus, given different assessments about the strategic environment, contradictory policy recommendations can be derived. In other words, while Lawson may not find the policy implications of Quackenbush's application to be sound, his objection would not necessarily be with Perfect Deterrence Theory. Rather, the source of his judgment would likely lie in a rejection of the initial conditions assumed by Quackenbush.[12]

For example, two economists, using the same theoretical construct may reach policy conclusions that are at odds with one another simply because one assumes a future low national economic growth rate and the other a much higher rate. In such cases, it is not the model or theory that leads to the policy disagreement; it is a lack of consensus about what the state of the world is or is going to be at the time the theory is applied. To reject or dismiss the theory, simply because of the discrepant policy recommendations, is to throw the baby out with the bath-water.[13]

Another reason for Professor Lawson's premature and unfounded rejection of Perfect Deterrence Theory may be traced to a fundamental misunderstanding of game theory in general and of its rationality postulate in particular. Since Perfect Deterrence Theory is comprised of a series of logically consistent game-theoretic models and rests on the assumption that the actors (players) in these models are "rational," one might easily concur with Lawson that it is highly "rationalistic." Whether that is a good or bad trait, however, depends entirely on what one means by rational. At times Professor Lawson seems to equate the concept with a neoclassical economists' view or what Herbert Simon calls "procedural rationality."[14] This definition of rationality presumes a near omniscient decision-maker who makes a "cool and clearheaded ends-means calculation"[15] that takes account of all possible courses of action and weights the pros and cons of each of them before making a decision. While this definition of rationality has certain merits,[16] it is

not the way rationality is defined or used in either game theory or in Perfect Deterrence Theory.

To be sure, there are some outliers, but most game theorists, including Quackenbush and myself, take an instrumentalist view of rationality, as Lawson clearly notes. Instrumental rationality involves a simple ends-means connection: players are rational when they take actions that they perceive will further their goals. Lawson is certainly correct when he notes the circularity of this definition; it is indeed a tautology, albeit a most useful one. The innocuous assumption of instrumental rationality allows an analyst to connect choices or decisions with policy preferences or goals. In other words, it simply presumes that decision-makers are purposeful, that when they act, they do so for a reason, and that they do not act randomly. These goals might be wrong-headed, they might be based on misinformation, they might be emotionally driven, and they might even be disastrous. But the players who pursue them are nonetheless "rational" *if* one takes a limited view of what rationality requires. Thus, unlike procedurally rational actors, instrumentally rational actors sound a great deal like ordinary, fallible, human beings. All of which is to say that the definition of instrumental rationality is more than consistent with Lawson's observation that "what is rational may well reflect the circumstances in which a decision is taken."[17] In his rush to condemn, Lawson fails to recognize that the very definition of rationality he rejects, is his own.

It should also be noted that the definition of procedural rationality presumes instrumental rationality, but not the other way around. In other words, the assumption of instrumental rationality does not rule out (or in) the "nuanced and imaginative thought processes…[that]… may well reflect the circumstances in which decisions are taken: constraints of time, information or resources; or the primary objectives at hand" that Lawson at once embraces and associates with Brodie, and his other theoretical role models.[18] All of which is to say that the assumption of instrumental rationality is a minimalist view that undermines Lawson's claim that Perfect Deterrence Theory is highly "rationalistic."[19] If anything, it presumes very little about human behaviour, which is one reason why game-theorists use it and one reason why it is so powerful an assumption. It is, therefore, more than ironic that Lawson at once denigrates the usefulness of this particular rationality postulate, wrongly associates it with the work of Wohlstetter, Kissinger and others, and then unwittingly uses it himself, throughout his essay, to draw conclusions about deterrence, as do most diplomatic historians and strategic theorists.[20] In the end, it is difficult to argue with Lawson's main conclusion that "fundamental misconceptions still haunt the study of deterrence."

Unfortunately, "Back to the Future" is a case in point. The linear relationship Lawson posits between Perfect Deterrence Theory and those who argue for missile defence systems and the wartime use of nuclear weapons simply does not exist. If anything, it actually runs the other way. For example, Lawson rightly attributes to Brodie and presumable other like-minded classical deterrence theorists, the view that the "risk of nuclear retaliation, following an attack, would be sufficient to convince an aggressor (or 'challenger') to refrain from carrying out the initial strike."[21] There is nothing in this statement that is inconsistent with the gestalt of Perfect Deterrence Theory.

Lawson's misconceptions, however, are not limited to a non-existent connection between the advocates of defence and proponents of Perfect Deterrence Theory. His understanding of the nature of game-theoretic models also leaves much to be desired. To wit: Lawson finds a simple model of deterrence that Quackenbush describes to be "problematic" for two reasons. First "it looks much more like a summary of the operation of defence than it does deterrence."[22] He says this simply because the model contains a *possible* choice for a defender of the status quo (or target). Why this is a problem, remains unclear. As Lawson himself asserts "[l]eaders have to decide all the time whether or not to challenge other countries. They no doubt do so on the basis of careful calculation about the response of the prospective target."[23] The rudimentary model Lawson finds problematic was designed to capture the very calculation that he claims is made "all the time." In other words, it is factually incorrect to claim as Lawson does that the model "ignores any threats of retaliation by the defender that precede the attacker's decision to strike."[24] The model simply takes Defender's threat as a given. Challenger's initial (node 1) decision is determined by its evaluation of the threat's potency (i.e., capability) and credibility (i.e., believability or rationality). Contrary to Lawson's assertion, the model sheds a great deal of "light on the mechanics of deterrence."[25] Under complete information, deterrence will succeed (no attack will take place) if Defender's threat is both capable and credible;[26] and it will fail if Defender's threat lacks either credibility or capability or both.[27]

Professor Lawson also finds fault with the model's implication that "deterrence is more likely to succeed if the defender prefers war to backing down in the face of an attack," a preference configuration that he correctly associates with a Prisoners' Dilemma game.[28] Conversely, when Defender prefers backing down to war, the model suggests that deterrence will fail. Again, Lawson rightly associates this preference configuration with the game of Chicken. He wrongly concludes, however, that the model is problematic because "applications of game theory to in-

ternational conflict generally demonstrate that Chicken games produce mutual cooperation more reliable than Prisoner's Dilemma games do."[29] The only source he gives for this claim is Snyder and Diesing's "exhaustive analysis of international crises."[30]

While there are many things to be admired in this classic, but now dated, study of interstate conflict, Snyder and Diesing's application of game theory to crisis decision-making is not one of them.[31] Perhaps Professor Lawson was led astray because he has relied on a source that is over thirty-five years old. Whatever the reason, his erroneous analysis rests on a misunderstanding about the definition of a game.

Games are defined by both their rules *and* the preference (pay-off) functions of the players. Thus games with different rules, but congruent preferences, constitute different games and may imply different outcomes.[32] Prisoners' Dilemma and Chicken are normal-form games in which the players are assumed to make a single strategy choice *before* the game begins. By contrast, the game models that constitute Perfect Deterrence Theory are extensive-form games in which the players make choices sequentially and (in games of incomplete information) update their beliefs about each other's preferences as they observe each other's action choices. Thus, the theoretical expectations of these models are quite different than those of the two classic games of Prisoners' Dilemma and Chicken, even when the preferences of the players in both sets of models are identical.

But even if this were not the case, Lawson's observations about the implications of Snyder and Diesing's empirical conclusions are also off the mark. Ironically, Kilgour and I point to these findings in our 2000 book as support for one of Perfect Deterrence Theory's key conclusions: successful deterrence generally requires threats that are not only capable, but credible as well.[33] By our definitions, both players in Chicken lack credible threats, but both players in a game with Prisoners' Dilemma-*like* preferences, have threats that are rational/believable/credible. Snyder and Diesing's case summary clearly suggests that deterrence is unlikely to succeed in games in which the players' preferences are the same as those of the players in Chicken, and likely to succeed when they are the same as those in Prisoners' Dilemma. Their empirical summary is patently more than consistent with the theoretical expectations of Perfect Deterrence Theory—unless one somehow concludes, as does Lawson, that the compromise outcome that Snyder and Diesing associate with those crises with the structural characteristics of a Prisoners' Dilemma game has nothing to do with deterrence. The compromise outcome in this game can only come about when both players *co-operate* with one another. It is not only my view, but it is also the view of Snyder and Diesing and

the vast majority of strategic analysts who use game theory to analyse interstate conflict, that mutual deterrence and mutual cooperation are one and the same thing.[34]

Professor Lawson also finds fault with Perfect Deterrence Theory's seemingly straightforward conclusion that one-sided (or unilateral) deterrence is easier to achieve than two-sided (or mutual) deterrence—a conclusion that has strong empirical support.[35] Lawson's problem stems from his belief, which is also consistent with the corpus of Perfect Deterrence Theory, that a power without nuclear weapons (i.e., without a capable threat) will not be able to deter a nuclear power dissatisfied with the status quo. Perfect Deterrence Theory's conclusions about the relative stability of unilateral and mutual deterrence relationships, however, rests on the condition that each actor in either type of deterrence relationship possess a threat that, when executed, will hurt the other. Thus, Lawson's ill-informed objection is beside the point. Moreover, Lawson's suggestion that Quackenbush somehow implied that classical deterrence theorists have claimed that "two aggressors would have a hard time trying to deter one another" is factually incorrect.[36] The inference exists only in Lawson's mind. In fact, Perfect Deterrence Theory explicitly rejects the argument of the vast majority of deterrence theorists who claim that parity constitutes a sufficient condition for deterrence success.[37] It is unfortunate indeed that the author of "Back to the Future" has chosen to disparage a theory he does not understand and a methodology he does not fully comprehend to defend a deficient and discredited theory that he firmly believes in.

In his rush to judgment, Professor Lawson also makes a specious connection between a general *relationship* of mutual deterrence and the *condition* of mutual assured destruction (MAD), which may (or may not) characterize any deterrence relationship, unilateral or mutual. Mutual assured destruction refers to a situation in which each of two states is unable to prevent an unacceptable retaliatory attack by the other. But in Perfect Deterrence Theory, a relationship of mutual deterrence is narrowly defined as one in which both states are dissatisfied with the status quo and, therefore, would prefer to alter it. (In a relationship of unilateral deterrence, there is only one dissatisfied state, the challenger). It should be clear that the vulnerability of one state to a retaliatory attack by another is independent of its satisfaction or dissatisfaction with the status quo. Were this not the case, the United States and Canada would currently be in a relationship of unilateral deterrence at best, and of mutual deterrence at worst.

Theories like Perfect Deterrence Theory are complexes of assumptions, definitions, and logically connected hypotheses.[38] The definitions

matter. Had Professor Lawson made an effort to understand how mutual deterrence is defined in Perfect Deterrence Theory, he would have likely not conflated two unrelated terms simply because they shared the same first name. Nowhere is his inattention to the specifics of the theory he attempts to diminish more obvious than when he recommends that deterrence theorists should "experiment with modeling threats of punishment and the ways that promises to retaliate might affect the initial decision whether or not to launch an attack"—for that is precisely what Perfect Deterrence Theory does.[39]

To summarize, Professor Lawson is certainly correct that Perfect Deterrence Theory can be thought of as a theoretical alternative to classical formulations of deterrence, Walt notwithstanding. Its axiomatic base, its theoretical characteristics, its major propositions, and its common-sense policy recommendations[40] clearly set it apart from the standard theory that traces its roots to the work of Bernard Brodie, Thomas Schelling, Glenn Snyder, and several others.[41] Nonetheless, it does not necessarily follow that Perfect Deterrence Theory's intellectual roots can be found in the work of those defence intellectuals whose policy recommendations included the wartime use of nuclear weapons. In point of fact, Perfect Deterrence Theory stands on its own, apart from the debates of the 1950s and early 1960s. And, in a way, it subsumes both of these approaches. The models that comprise Perfect Deterrence Theory allow for an examination of a variety of strategic circumstances that may vary according to the credibility of threats, the capability of threats, and the satisfaction of the players. What those conditions are at any one point in time is an empirical rather than a theoretical question. Thus, Lawson could make his own assessment of the strategic environment, and then use Perfect Deterrence Theory to determine the logical consequences of that assessment. What may either surprise or dismay Professor Lawson is the strong possibility of a convergence between many of his expectations and those of a theory he has not made a good-faith effort to scrutinize. ■

Frank C. Zagare is UB Distinguished Professor of Political Science at the University at Buffalo, SUNY. His main research interests lie in the nexus between security studies and game theory. Professor Zagare's theoretical work has focused on deterrence, crises, conflict escalation, and bargaining and negotiation. He is the author, co-author, or editor of six books including The Games of July, Perfect Deterrence (with D. Marc Kilgour), The Dynamics of Deterrence, and Game Theory: Concepts and Applications.

[1] Frank C. Zagare, "Classical Deterrence Theory: A Critical Assessment," *International Interactions* 21, no. 4 (1996): 365-387; Frank C. Zagare, "Reconciling Rationality with Deterrence: A Re-examination of the Logical Foundations of Deterrence Theory," *Journal of Theoretical Politics* 16, no. 2 (2004): 107-41.

[2] Frank C. Zagare and D. Marc Kilgour, *Perfect Deterrence* (Cambridge: Cambridge University Press, 2000); Frank C. Zagare, *The Games of July: Explaining the Great War* (Ann Arbor: University of Michigan Press, 2011).

[3] Stephen L. Quackenbush, "National Missile Defense and Deterrence," *Political Research Quarterly* 59, no. 4 (2006): 533-41.

[4] Stephen L. Quackenbush, "Deterrence Theory: Where Do We Stand?" *Review of International Studies* 37 (2011): 762.

[5] Michael Charlton, *From Deterrence to Defense: The Inside Story of Strategic Policy* (Cambridge, MA: Harvard University Press, 1987).

[6] Stephen M. Walt, "Rigor or Rigor Mortis? Rational Choice and Security Studies," *International Security* 23, no. 4 (1999): 5.

[7] The assumption of undifferentiated actors is a modal assumption of modern realist theory of which classical deterrence theory is a variant. For this argument, see Jeffrey W. Legro and Andrew Moravcsik, "Is Anybody Still a Realist," *International Security* 24, no. 2 (1999): 5-55. For an example, see Kenneth N. Waltz, *Theory of International Politics* (Reading, MA: Addison-Wesley, 1979).

[8] Zagare, "Reconciling Rationality," 135. After this was pointed out in an initial review of "Back to the Future," Lawson added a footnote acknowledging the differences between my conclusions about the effectiveness of a national missile defence system and Quackenbush's. This significant fact, however, did not deter him from clinging to an argument that falls apart once it is established.

[9] It is difficult to understand how this highly qualified statement can be construed as an endorsement of either the development or the deployment of a national missile defence system.

[10] Quackenbush, "National Missile Defense" also notes that the dissatisfaction with the status quo that results from the deployment of a defensive system may undermine strategic stability.

[11] Quackenbush, "Deterrence Theory," 762.

[12] None of this should be taken as a rejection of Quackenbush's argument which stands on its own.

[13] Of course, there may be other reasons for rejecting the theory. Logical inconsistency and/or empirical inaccuracy stand out as obvious grounds for dismissal.

[14] Herbert A. Simon, "From Substantive to Procedural Rationality," in *Method and Appraisal in Economics*, ed. S.J. Latsis (Cambridge: Cambridge University Press, 1976).

[15] Sidney Verba, "Assumptions of Rationality and Non-rationality in Models of International Systems," in *The International System: Theoretical Essays*, eds. Klaus Knorr and Sidney Verba (Princeton: Princeton University Press, 1961): 95.

[16] Frank C. Zagare, "Rationality and Deterrence," *World Politics* 42, no. 2 (1990): 238-260.

[17] Fred Lawson, "Back to the Future in the Study of Deterrence", *St Antony's International Review* 9 no. 1 (2013): 149.

[18] Ibid., 149.

[19] It is, however, "abstract" as Lawson charges. But all theories, by their nature, are abstractions.

[20] For example, Lawson's assertion that deterrence will tend to prevail "once both states gain the capability to inflict certain destruction on one another" rests on the innocuous assumption (of instrumental rationality) that both states prefer not to be destroyed and that they act accordingly.

[21] Lawson, "Back to the Future", 145.

[22] Ibid., 146.

[23] Ibid., 146.

[24] Ibid., 147.

[25] Ibid., 147.

[26] When deterrence succeeds, Challenger is never faced with a choice (at node 2).

[27] The model's implications are slightly different when information is incomplete. Complete and incomplete information are technical terms in game theory. In a game of complete information, the players know each other's preferences (pay-off functions). If that is not the case, the game is said to be a game of incomplete information.

[28] Lawson, "Back to the Future", 148.

[29] Ibid., 148.

[30] Glenn H. Snyder and Paul Diesing, *Conflict among Nations: Bargaining, Decision Making and System Structure in International Crises* (Princeton: Princeton University Press, 1977).

[31] For example, Snyder and Diesing neither mention nor use the concept of a Nash equilibrium, the accepted measure of rational behaviour and the basis of all solution concepts for the normal- (or strategic-) form games they explore.

[32] Lawson seems to recognize this, but only when discussing the differences between a simultaneous choice and a sequential choice play of Chicken.

[33] Zagare and Kilgour, *Perfect Deterrence*, 299.

[34] Snyder and Diesing, *Conflict among Nations*, 497. See also Robert Jervis, "Deterrence Theory Revisited," *World Politics* 31, no. 2 (1979): 289-324 .

[35] Paul D. Senese and Stephen L. Quackenbush "Sowing the Seeds of Conflict: The

Effect of Dispute Settlements on Durations of Peace," *Journal of Politics* 65, no. 3 (2003): 696–717; Stephen L. Quackenbush and Jerome F. Venteicher II, "Settlements, Outcomes, and the Recurrence of Conflict," *Journal of Peace Research* 45, no. 6 (2008): 723–42.

[36] Lawson, "Back to the Future", 151.

[37] Zagare, *Games of July*, 42.

[38] Lisa Martin, "The Contributions of Rational Choice: A Defense of Pluralism," *International Security* 24, no. 2 (1999): 74-83.

[39] Lawson, "Back to the Future", 153.

[40] Perfect Deterrence Theory provides theoretical support for a policy of minimum deterrence and for meaningful arms control agreements. It also recommends conditionally cooperative foreign policy initiatives and negotiating stances that are based on reciprocity. Conversely, it opposes both an "overkill" capability and the proliferation of nuclear and other weapons of mass destruction.

[41] For a concise comparison, see Zagare, *Games of July*, 42.