**Methodology as ideology: the prisoner's dilemma as misapplied to trench warfare**

Andrew Gelman[1]
Department of Statistics and Department of Political Science, Columbia University, New York

**Abstract**

*The Evolution of Cooperation*, by Axelrod (1984), is a highly influential study that identifies the benefits of cooperative strategies in the iterated prisoner's dilemma. We argue that the most extensive historical analysis in the book, a study of cooperative behavior in First World War trenches, is in error. Contrary to Axelrod's claims, the soldiers in the Western Front were not in a prisoner's dilemma (iterated or otherwise), and their cooperative behavior can be explained much more parsimoniously as immediately reducing their risks. We discuss the political implications of this misapplication of game theory.

Keywords: cooperation, First World War, game theory, prisoner's dilemma

[1] We thank Hayward Alker for his guidance throughout this project and Charles Cameron for helpful discussions.

**Methodology as ideology: the prisoner's dilemma as misapplied to trench warfare**


## Abstract

*The Evolution of Cooperation*, by Axelrod (1984), is a highly influential study that identifies the benefits of cooperative strategies in the iterated prisoner's dilemma. We argue that the most extensive historical analysis in the book, a study of cooperative behavior in First World War trenches, is in error. Contrary to Axelrod's claims, the soldiers in the Western Front were not in a prisoner's dilemma (iterated or otherwise), and their cooperative behavior can be explained much more parsimoniously as immediately reducing their risks. We discuss the political implications of this misapplication of game theory.



Keywords: cooperation, First World War, game theory, prisoner's dilemma


## Introduction

Robert Axelrod's *The Evolution of Cooperation* (1984) is an extremely influential work in the application of game theory to political science. The book uses theory, computer experiments, and historical examples to identify the benefits of cooperative strategies in the iterated prisoner's dilemma. The book's longest historical example is a model of cooperative behavior among British and German soldiers in First World War trenches. The model is appealing both intellectually, as a solution to a behavioral puzzle, and emotionally, because it seems to bring some sense to a scary and confusing subject. However, a closer look reveals serious problems with the application of the prisoner's dilemma to this particular historical situation, which in turn calls into question the claims made about the importance of the prisoner's dilemma in studying cooperation in general.

The fundamental problem Axelrod was studying is *cooperation.* From a psychological or economic point of view, why do people cooperate with each other

(instead of acting purely selfishly, which would give them short-term benefits, at the very least)? From a game-theoretic standpoint, cooperation has always been viewed as a puzzle and has been given names such as the "free rider problem," the "prisoner's dilemma," and the "tragedy of the commons" (Hardin, 1967).

The key question in settings described by this mathematical formulation is, why do people cooperate at all in situations like this, or, to reframe it more constructively, how can we construct social systems to encourage cooperation? In the long term, cooperation makes sense, but in the short term, it pays to not cooperate. How can cooperative short-term behavior occur? Several answers have been suggested. Psychologically, people seem to feel more comfortable cooperating with people they know, and this has been studied experimentally by economists and psychologists (see Dawes, Kragt, and Orbell, 1988). In situations where cooperation is important (for example, in a business) or even a matter of life and death (for example, in the military), it is considered crucial to set up a "team spirit."

However, in other settings, most notably in the economic sphere, the incentives to not cooperate are so strong that psychological motivation does not seem enough. Cooperation can then be enforced through governmental action or private binding agreements (which themselves typically require governmental presence to be enforceable). These situations where cooperative behavior requires outside enforcement are called "externalities" in the terminology of economics.

Axelrod's interest was slightly different, however. Rather than study settings where cooperation is automatic, or where cooperation needed outside enforcement, he was interested in intermediate scenarios in which cooperative behavior was risky and unstable but developed anyway. This seems to describe much of human interactions—when the rules break down, people can act brutally, but stable societies are greased by a layer of trust.

In his book, Axelrod made a three-part argument. First, at the level of pure game theory, he argued—and presented evidence for the claim—that a strategy called "tit for tat" (TFT) was effective in a game consisting of repeated plays of the prisoner's dilemma. (The TFT strategy is defined as follows: begin by cooperation, and then at each step do what your opponent did in the previous step.) Second, Axelrod argued that

this strategy was effective in important real-life settings where these games arise. His central example was the behavior of soldiers in trench warfare in World War I (an example perhaps chosen because it is well-documented and was a relatively stable system for years, which would presumably allow things to settle into equilibrium states if there were any). The third step of Axelrod's argument was evolutionary: because TFT was effective, it would be natural for organisms that developed this strategy to survive and reproduce; hence, one would expect this strategy to prevail in nature. We do not discuss this third step further here.

**Axelrod's application of the prisoner's dilemma to trench warfare**

Axelrod looked at trench warfare in World War I, which, as detailed in the fascinating book by Tony Ashworth (1980), had ongoing examples of cooperative behavior amidst a violent, anticooperative structure. The soldiers in the two opposing armies can be considered as two players in a game, in which at each step a player can cooperate by *not* shooting at the other side. Axelrod's basic argument goes as follows: at each step, a player is better off if he does not cooperate (that is, if he shoots), since this has the chance of eliminating an enemy soldier. However, if both sides were somehow to agree to cooperate, then they would both be better off, since none of them would be shot. Empirically, the two sides *did* in fact cooperate despite the many obstacles placed in their path. We shall briefly describe the story of the trench-warfare cooperation (as related by Ashworth in great detail) and then return to the game-theoretic analysis.

World War I started in the summer of 1914 when the Germans invaded Belgium. By the late fall of that year, the war in the Western Front (France and Belgium) had stabilized and the two sides (Germans and their allies on one side, French and British and their allies on the other) were facing each other behind trenches. The soldiers were expected to shoot at each other from the trenches but in many instances avoided doing so, and Christmas, 1914, featured many instances of the two sides meeting peacefully between their front lines. As described by Ashworth, the commanders on both sides did not like this and ordered the soldiers to shoot at each other. At some point, this pattern

also switched to cooperation in many places, with soldiers shooting to miss on purpose (and at the same time demonstrating their ability to do harm by aiming at precise targets). Throughout the war, the commanders tried different strategies—for example, rotating troops more quickly in and out of the front lines—to stop the troops from getting friendly with the enemy. The most effective strategy appeared to be sending the soldiers on raids into the enemy trenches. This put them in a kill-or-be-killed situation in which cooperation was essentially impossible.

This pattern—soldiers who do not want to fight and commanders who force them to do so—has been reported throughout history, as has been noted by the popular military historian John Keegan (1976); for example, officers in the Napoleonic Wars stood behind their troops with bayonets to force them toward the enemy. In the Second World War, a famous study by Colonel S. L. A. Marshall (1947) estimated that only one-quarter of the U.S. soldiers in a position to fire their rifles actually did so (although this finding has been questioned; see Spiller, 1988). This behavior has been attributed to fear and a sense of isolation, as well as simple self-preservation, since firing your gun can make you a target.

Now we return to Axelrod's argument, which is an attempt to explain theoretically the cooperation described by Ashworth. Given the immediate risks of cooperation, how did the soldiers so many times develop cooperative social structures without the possibility of binding agreements? Axelrod suggests they were following the TFT strategy: starting by cooperating, and then continuing to cooperate as long as cooperation was followed on the other side. In the trench warfare example, this means: do not shoot until your opponent shoots. If your opponent shoots, shoot back, but then stop shooting if he stops.

In the terminology of game theory, TFT works because trench warfare is a *repeated-play* game. In a single play, the dominant strategy is non-cooperation—that is the essence of the tragedy of the commons or the prisoner's dilemma. But when the game is played repeatedly, it is beneficial to establish a pattern of cooperation, a point that Axelrod illustrated with a simulation in which, out of a large selection of game strategies, TFT performed best. The simulation was performed in an "evolutionary" way, with the more successful strategies "reproducing" to reflect the popularity of success, and TFT performed even better as time went on and the alternative, non-cooperative strategies

diminished in the population.

To summarize the argument: soldiers spontaneously developed cooperation strategies despite the short-term advantages of shooting at the enemy. This behavior is "explained" or at least can be modeled by a repeated-play prisoner's dilemma game, in which cooperative strategies will grow to dominate. With a large population of such players, the norm of cooperation becomes stable, to the extent that, in the First World War, commanders had to actually change the rules of the game (for example, by forcing soldiers to make raids) in order to eliminate cooperative behavior. This argument is appealing because it explains otherwise baffling behavior (how did the soldiers develop the norm of cooperation without the ability to directly communicate or make agreements) and also, from a moral point of view, it gives some hope that cooperation might be possible in a hostile world.

**Why trench warfare is actually *not* a prisoner's dilemma**

Having summarized Axelrod's argument, we now explain why we think it is mistaken. The model's key assumption is that an individual soldier benefits, in the short term, from firing at the enemy. (In the terminology of the prisoner's dilemma, to cooperate is to avoid firing, and the model assumes that, whatever the soldiers on the other side do, you are better off firing, that is, not cooperating.) Thus, elaborate game-theoretic modeling is needed to understand why this optimal short-term behavior is not followed. In fact, however, it seems more reasonable to suppose that, as a soldier in the trenches, you would do better to *avoid* firing: shooting your weapon exposes yourself as a possible target, and the enemy soldiers might very well shoot back at where your shot came from (a point mentioned by S. L. A. Marshall). If you have no short-term motivation to fire, then cooperation is completely natural and requires no special explanation. This is an example of a fancy mathematical model being created to explain behavior that is perfectly natural. In fact, if any games are being played, they are between the soldiers and the commanders on each side of the front line, with many of the soldiers avoiding combat and the commanders trying to enforce it.

Trench warfare is a setting in which both sides benefit in the immediate sense from cooperation, so that no subtle game-theoretic models are needed to explain it (although, as a practical necessity, some coordination is needed to establish the exact forms of cooperation, as described by Ashworth).

If the explanation of cooperative behavior in the trenches is indeed so obvious, then how could a researcher as clever and well-read as Axelrod get it so wrong (and how could his book have been so widely praised)? To start with, in any theoretical formulation there is a natural bias toward conventional roles. In game theory, it is usual to think of the two sides in a war as opponents. And, indeed, in the grand strategy of the First World War, this might be reasonable (although even this is complicated, since it has been argued that France lost the war by winning it). But on the front line, it is not at all clear that shooting an individual German would help a given British soldier.

Another reason why Axelrod might have been led to construct his complicated model is that, historically speaking, cooperation among soldiers on opposite sides of a battle has been unusual, and thus an elaborate explanation might seem to be required. However, this does not really address the question of whether his fundamental assumptions are reasonable. From the perspective of game theory, other models such as coordination games might be more appropriate (see, for example, Snidal, 1985). In fact, the very stability of cooperation in First World War trenches, and the fact that the commanders had to devise elaborate schemes to thwart it, argues for the claim that joint cooperative behavior was a stable solution to the "game" for soldiers on both sides, and thus it was no prisoner's dilemma at all.

**Political implications**

Does this matter? How does a difference in opinion about game theory affect our understanding of military or social behavior? In some ways, not at all—we are all trying to explain the same facts. But we would like to argue that the mathematical model and its interpretation do matter. As always, arguments about history are often really about the present and the future. If you really believe the prisoner's dilemma and tit-for-tat story,

then you are characterizing the soldiers on the two sides of the First World War as, fundamentally, opponents, who would only avoid shooting at each other because of long-term calculations based on repeated encounters. Looking at the situation more carefully, it appears that the soldiers on both sides had a common immediate interest in not fighting, hence the need for their commanders to develop tactics in which combat was unavoidable.

We conclude our discussion of this example with a question: what are the political implications of Axelrod's theory? At one level, the theory looks "liberal," both in the so-called classical or 19th-century sense of respecting individual action (that is, in contrast to "classical conservative" political theory which favors tradition) and in the more recent American sense of supporting altruistic behavior. The story goes as follows: it's a cold world out there, with payoff matrices that favor non-cooperation, but with multiple plays of the prisoner's dilemma game, if everyone acts in his long-term interest, virtue (in the sense of cooperation or, in the warfare setting, nonviolence) will be rewarded. This appears to be an intriguing synthesis of classical and modern liberal ideas.

However, in another sense, Axelrod's game-theoretic model is fundamentally conservative, because it accepts the assumption that the soldiers on the two sides of the trenches had immediate motivation to shoot at each other. His recommendation for cooperation and peacemaking is to accept existing conflicts and work within their structure, rather than to suggest, as a liberal might, that these conflicts only exist because the commanders of the armies are exercising unchecked power.

We don't mean to imply any claim about Axelrod's personal beliefs here. His book was, and perhaps continues to be, influential, and we suspect that one reason for its success is that it appears to supply a theoretical and empirical justification of the liberal idea of cooperation. Since at least the days of Machiavelli and continuing to the present time, political theory has often leaned toward the position that individuals and nations must pursue their own self-interests (see Miller, 1999, for a perspective from psychology). At this point, Axelrod's argument came as a welcome relief: a rigorous justification of altruistic behavior. But, at least in the trench warfare example, his logic is unnecessarily complicated, and in fact cooperation occurred for much more direct reasons.

## Conclusions

In summary, some carefully documented history of the First World War revealed the surprising fact that soldiers on the Western Front routinely avoided fighting. This finding inspired quantitative modeling. The evolutionary repeated-play formulation of the prisoner's dilemma is a model—a logical story expressed in mathematics—showing how cooperative behavior can arise and thrive in an initially hostile environment. The model is interesting in their own right, but before applying them to warfare or other conflict settings, one should check whether its basic premise applies. If cooperation (not fighting) has short-term benefits—as it often does—then the prisoner's dilemma does not apply.

Looking toward the future, how can we achieve more cooperation in real-life "tragedies of the commons" such as environmental devastation and international arms races? Axelrod's logic suggests that we should look for ways to set these up as repeated-play games so that participants have long-term motivations to cooperate. A more skeptical approach (suggested by the analysis in this paper) might be to set up immediate gains from cooperation and to watch out for outside agents (such as the commanders in the trench warfare example) who have motive and opportunity to disrupt the cooperative behavior.

**References**

Ashworth, T. (1980). *Trench Warfare, 1914-1918: The Live and Let Live System.* New York: Holmes and Meier.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Dawes, R., Kragt, A. J. C., and Orbell, J. M. (1988). Not me or thee but we: the importance of group identity in eliciting cooperation in dilemma situations: experimental manipulations. *Acta Psychologica* **68**, 83-97.

Hardin, G. R. (1968). The tragedy of the commons. *Science* **162**, 1243-1248.

Keegan, J. (1976). *The Face of Battle*. London: Penguin.

Marshall, S. L. A. (1947). *Men Against Fire.* Washington, D.C.: The Infantry Journal, and New York: William Morrow.

Miller, D. T. (1999). The norm of self-interest. *American Psychologist* **54**, 1053-1060.

Snidal, D. (1985). Coordination versus prisoners' dilemma: implications for international cooperation and regimes. *American Political Science Review* **79**, 923-942.

Spiller, R. J. (1988). S. L. A. Marshall and the ratio of fire. *RUSI Journal* **133**, 63—71.