The Game Theory of Mutually Assured Destruction

Pranay Gundam

June 4, 2024

Introduction

The game theory we usually talk about in introductory/intermediate undergrad classes and high-school always left me really dissatisfied (which to be fair, it is my fault that this is the extent of classes that I have taken that cover game theory). I know there are textbooks that cover models and games that are a quintessential part of the literature but I wanted address modeling an interaction by myself. Specifically I want to talk about the idea of Mutually Assured Destruction (MAD) that became so popular during the Cold War ...

The Basic 2x2

In highschool AP Econ classes we are taught about simple games where there are two players each of whom can take one of two possible actions. In the context of MAD, we could label the two agents as the "US" and "Russia", both of whom can decide to "launch nukes" or "don't launch nukes". The corresponding chart we would use to work this problem out would look something like

I made the chart above with a set of payoffs for each combination of actions such that there is only one Nash equilibrium at both agents choosing "Don't Launch". The equilibria of this game could have changed if I had chosen different payoffs and one interesting concept to explore is the conditions on the payoffs in order for each outcome in the statespace to become a Nash equilibrium. Specifically, consider the chart below, here the x payoffs belong to the US and the y payoffs belong to Russia. The $x_{DL,DL}$ payoff for example is the payoff that the US experiences when both the US and Russia chooses not to launch.

Visualizing the domains at which these conditions hold is a bit difficult to do all at once since we have to consider a four dimensional space but we can at least talk about it.

We can also simplify the problem of visualizing when certain outcomes become Nash equilibria by imposing some functional constraints on the payoffs. For example, we can say that a country

has a utilitarian mindset applying to people of all nationalities (where all life has at least some positive value) and when getting launched at it is weakly more utility for them to not retaliate. In such a case, if the US were to adopt this mindset, we would then have that $x_{DL,L} \geq x_{L,L}$. This combined with the not so wild assumption that the US would prefer not to get launched at would yield the relationship $x_{DL,DL} \geq x_{DL,L} \geq x_{L,L}$. We can tack on one final assumption, that the US attributes more value to their own citizens than citizens of other countries, and we can completely describe the relationship between all the payoffs that the US experiences

$$x_{DL,DL} \geq x_{L,DL} \geq x_{DL,L} \geq x_{L,L}.$$

What is interesting to discuss is how countries/agents with different functional paradigms of determining their payoffs fare in contest with eachother.

N Discrete Choices