CONTACTPROJECT.org

IS HUMANITY READY FOR CONTACT WITH EXTRATERRESTRIAL INTELLIGENCE?

Dark Forest Game Theory Equations

Erich Habich-Traut

June 2025

Let's formulate some equations to represent the scenarios described in the Dark Forest Hypothesis, using a game theory framework. We'll define the payoffs for two civilizations, C_1 and C_2, based on their chosen strategies. Let $P(S_1, S_2)$ be the payoff for C_1 when C_1 chooses strategy S_1 and C_2 chooses strategy S_2. Similarly, $P(S_2, S_1)$ will be the payoff for C_2.

Let the strategies be:

- H: Hide (Stay Silent)
- B: Broadcast (Reveal Yourself)
- A: Attack First (Preemptive Strike)

We'll use variables to represent the utilities (or disutilities) of different outcomes.

Assumptions for Initial Dark Forest Model:

- $U_{survive}$: Utility of long-term survival. (This is the baseline and what players aim to maximize).
- C_{attack}: Cost of launching an attack (resource expenditure).
- $D_{annihilation}$: Disutility of being annihilated.
- $B_{cooperation}$: Benefit of cooperation and shared knowledge.
- $R_{discovery}$: Risk (disutility) of being discovered later when the other civilization is stronger.
- S_{fail_attack}: Disutility of a failed attack leading to retaliation.

—

Payoff Matrix for a Single Interaction (Dark Forest - Simplified)

This matrix represents the perspective of Civilization 1. The payoff for Civilization 2 would be symmetrical if they are identical in their utility functions. Let's assume a zero-sum or largely competitive game for the initial Dark Forest model.

C_1 C_2	**Hide (H)**	**Broadcast (B)**	**Attack (A)**
Hide (H)	$(U_{survive} - R_{discovery})$ for C_1, $(U_{survive} - R_{discovery})$ for C_2	$(D_{annihilation})$ if C_2 is hostile, $(U_{survive} + B_{cooperation})$ if C_2 is cooperative	$(D_{annihilation})$
Broadcast (B)	$(D_{annihilation})$ if C_2 is hostile, $(U_{survive} + B_{cooperation})$ if C_2 is cooperative	$(U_{survive} + B_{cooperation})$ if cooperative, $(D_{annihilation})$ if hostile	$(D_{annihilation})$
Attack (A)	$(U_{survive} - C_{attack})$ (successful attack)	$(U_{survive} - C_{attack})$ (successful attack)	$(U_{survive} - C_{attack})$ (successful attack) if C_1 wins, $(D_{annihilation})$ if C_2 wins

Explanation of Payoffs (Initial Dark Forest Assumptions):

- **H vs. H:** Both hide. They survive for now, but there's a risk of being discovered later by a stronger opponent.
- **H vs. B:** If C_1 hides and C_2 broadcasts:
  - If C_2 is hostile, C_1 is annihilated because it revealed itself.
  - If C_2 is cooperative, C_1 could benefit (though C_1 chose to hide, C_2 might reach out). This is a bit ambiguous for a strict Dark Forest model, where cooperation is less likely. For simplicity in the "Attack First" logic, we assume cooperation is rare or risky.
- **H vs. A:** If C_1 hides and C_2 attacks: C_1 is annihilated.
- **B vs. H:** Symmetrical to H vs. B.
- **B vs. B:** Both broadcast.
  - If cooperative: Mutual benefit.
  - If hostile: Mutual annihilation (or one annihilates the other). The Dark Forest implies the latter.
- **B vs. A:** If C_1 broadcasts and C_2 attacks: C_1 is annihilated.
- **A vs. H:** If C_1 attacks and C_2 hides: C_1 successfully eliminates the threat, paying the cost of attack.
- **A vs. B:** If C_1 attacks and C_2 broadcasts: C_1 successfully eliminates the threat.
- **A vs. A:** Both attack. One wins, one loses. For simplicity, we can assume C_1 wins with some probability p, or it's a mutual annihilation. In the "Attack First" logic, it's about minimizing the risk of being attacked.

—

Simplified Payoff Matrix focusing on the "Attack First" Dominant Strategy:

Let's assign numerical values to make the "dominant strategy" clear, assuming $D_{annihilation}$ is extremely negative. We'll simplify to a game where being annihilated is the worst outcome. Consider the "risk of annihilation" as the primary driver.

- Let $A = AnnihilationPayoff \approx -\infty$
- Let $S = SurvivalPayoff > 0$
- Let $C = CostofAttack > 0$
- Let $R = RiskofDiscoveryLater > 0$ (a small negative impact on survival)
- Let $B = BenefitofCooperation > 0$

C_1 C_2	**Hide (H)**	**Broadcast (B)**	**Attack (A)**
Hide (H)	$(S-R, S-R)$	(A, A) if hostile; $(S+B, S+B)$ if cooperative	$(A, S-C)$
Broadcast (B)	(A, A) if hostile; $(S+B, S+B)$ if cooperative	(A, A) if hostile; $(S+B, S+B)$ if cooperative	$(A, S-C)$
Attack (A)	$(S-C, A)$	$(S-C, A)$	$(S-C, S-C)$ (assuming successful attack for C_1)

—

Dominant Strategy Analysis (Attack First):

From C_1's perspective:

- **If C_2 Hides (H):**
  - C_1 Hides: $S-R$
  - C_1 Attacks: $S-C$

  If $S-C > S-R$, then Attack is better. This holds if $R > C$ (risk of discovery is greater than cost of attack).

- **If C_2 Broadcasts (B):**
  - C_1 Broadcasts: A (annihilation by hostile C_2) or $S+B$ (cooperation). Given the Dark Forest premise, A is highly likely.
  - C_1 Attacks: $S-C$

  $S-C > A$. So Attack is better.

- **If C_2 Attacks (A):**

- C_1 Hides: A
- C_1 Broadcasts: A
- C_1 Attacks: $S - C$

$S - C > A$. So Attack is better.

Under these assumptions, "Attack First" appears to be the dominant strategy because it's the only one that guarantees survival (albeit with a cost C) against a potentially hostile opponent, and avoids the near-certain annihilation from other strategies when the opponent is aggressive.

—

Flaws in the Dark Forest Game Theory - Incorporating New Variables

1. Mutually Assured Destruction (MAD):

Let's introduce a probability of successful attack, $p_{success}$. And a disutility for failed attack leading to retaliation, $D_{retaliation} \ll 0$. Now, the payoff for "Attack (A)" changes:

If C_1 attacks C_2:

- With probability $p_{success}$, C_1 gets $(U_{survive} - C_{attack})$ and C_2 gets $D_{annihilation}$.
- With probability $(1 - p_{success})$, C_1 gets $D_{retaliation}$ (or worse, $D_{annihilation}$ if C_2 successfully retaliates) and C_2 survives to retaliate.

The expected utility of attacking becomes: $E[U_{attack}] = p_{success}(U_{survive} - C_{attack}) + (1 - p_{success})D_{retaliation}$ If $p_{success}$ is low, or $D_{retaliation}$ is extremely negative (as in MAD), then $E[U_{attack}]$ can become very low, possibly lower than hiding.

2. Detection is Inevitable:

If hiding is impossible, then the H strategy effectively becomes equivalent to B in terms of detection. The "Pros" of hiding disappear. Let $R_{detection\_inevitable}$ be the disutility of being detected when hiding. If this value approaches $D_{annihilation}$, then:

- Payoff of (H, H) $\rightarrow D_{annihilation}$
- Payoff of (H, B) $\rightarrow D_{annihilation}$
- Payoff of (H, A) $\rightarrow D_{annihilation}$

This effectively removes "Hide" as a viable strategy for survival, pushing the game towards Broadcast or Attack.

3. Not All Civilizations Are Rational Killers:

This introduces different "types" of players. Let $P_{hostile}$ be the probability that a civilization is hostile, and $P_{cooperative}$ be the probability it is cooperative. The expected payoff of broadcasting (B) now depends on the opponent's type: $E[U_{broadcast}] = P_{hostile} \cdot D_{annihilation} + P_{cooperative} \cdot (U_{survive} + B_{cooperation})$ If $P_{cooperative}$ is high enough, and $B_{cooperation}$ is significant, then $E[U_{broadcast}]$ could outweigh the expected utility of attacking.

—

Alternative Equilibrium: The "Quiet but Armed" Strategy

Let Q: Quiet but Armed (Hide and Build Defenses) Let $C_{defense}$: Cost of building defenses. Let $D_{deterrence}$: Disutility for an attacker if they face strong defenses. (This lowers the expected payoff of attacking your civilization).

New Strategy: Quiet but Armed (Q)

C_1 C_2	**Quiet (Q)**	**Broadcast (B)**	**Attack (A)**
Quiet (Q)	$(U_{survive} - C_{defense}, U_{survive} - C_{defense})$ (Cold War Stalemate)	$(D_{annihilation})$ if C_2 is hostile; $(U_{survive} + B_{cooperation} - C_{defense})$ if cooperative (but C_1 is quiet)	$(D_{annihilation})$ if attack succeeds; $(U_{survive} - C_{defense} - S_{fail\_attack})$ if C_1 retaliates successfully
Broadcast (B)	(Symmetrical to Q vs B)	(Same as before)	(Same as before)
Attack (A)	$(p'_{success}(U_{survive} - C_{attack}), (1 - p'_{success})D_{retaliation})$ where $p'_{success} < p_{success}$ due to C_2's defenses	(Same as before)	(Same as before)

In the "Quiet but Armed" scenario, the probability of a successful attack against a "Quiet" civilization ($p'_{success}$) is lower than against a "Hiding" or "Broadcasting" one. This increases the attacker's $D_{retaliation}$ risk and reduces their expected payoff, leading to deterrence.

Nash Equilibrium in "Quiet but Armed":

If both players choose "Quiet (Q)", and the cost of defense is less than the expected cost of an attack or annihilation, and the deterrence is effective, then (Q, Q) could be a stable Nash Equilibrium. The condition for this equilibrium would be: $U_{survive} - C_{defense} > E[U_{attack}]$ $U_{survive} - C_{defense} > E[U_{broadcast}]$ (if C_2 is cooperative, but C_1 remains quiet) This framework allows for the exploration of various scenarios and the conditions under which different outcomes (annihilation, cooperation, cold war) might prevail in the cosmic arena.