Project status update: Strategic Adaptation in the "Gun Game" under Partial Observability

Course: Multi-Agent AI Systems

January 21, 2026

Project Team Members

ID	Degree	Name	Role
318317526	M.Sc.	Tom Badash	Theory Lead
209196104	B.Sc.	Rom Sheynis	Development Lead
316352715	B.Sc.	Lioz Shor	Analysis and Research Lead

1 Overview

In this project status update, we present the theoretical analysis baseline that underpins the remainder of our work. This baseline establishes (i) a formal model of the game and information structure, (ii) a belief-maintenance mechanism under partial observability, and (iii) a value-based method for deriving belief-dependent policies. Together, these components set the environment for the next stages of implementation, simulation, and persona-based evaluation.

Our research focuses on a finite-horizon, simultaneous-move game called the *Gun-Wall Game*. The Gun-Wall Game is an adaptation of the original *Gun Game* (also known as "007"), which follows the same core rules (Shoot, Reload, Block) but assumes full observability of the opponent's actions and ammunition. In contrast, in the Gun-Wall Game, players observe only round outcomes and must act under partial information.

2 Game definition

2.1 State space and action constraints

The game proceeds over a finite horizon of $T = 5$ rounds, indexed by $t \in \{1, 2, 3, 4, 5\}$. Each player $i \in \{1, 2\}$ holds binary ammunition:

$$a_t^i \in \{0, 1\}, \qquad x_t = (a_t^1, a_t^2) \in \{0, 1\}^2.$$

Thus, there are four underlying states:

$$(0,0),\ (1,0),\ (1,1),\ (0,1).$$

Players choose actions simultaneously each round. Let u_t^i denote player i's action at time t, and let the joint action be (u_t^1, u_t^2). The action set is

$$A = \{S, B, R\},$$

where S is *Shoot*, B is *Block*, and R is *Reload*. Legal actions depend on the player's own ammo:

$$A_i(a_t^i) = \begin{cases} \{R, B\}, & a_t^i = 0, \\ \{S, B\}, & a_t^i = 1. \end{cases}$$

State x_t	**Meaning**	**P1 legal actions**	**P2 legal actions**
$(0,0)$	both unarmed	$\{R, B\}$	$\{R, B\}$
$(1,0)$	P1 armed, P2 unarmed	$\{S, B\}$	$\{R, B\}$
$(1,1)$	both armed	$\{S, B\}$	$\{S, B\}$
$(0,1)$	P1 unarmed, P2 armed	$\{R, B\}$	$\{S, B\}$

2.2 Outcome signal under partial observability

Under partial observability, players do not observe opponent actions nor opponent ammo. Instead, after each round they observe an outcome signal:

$$o_t \in O, \qquad O = \{\text{Continue},\ \text{P1Win},\ \text{P2Win},\ \text{Tie}\}.$$

The outcome is a deterministic function of the underlying state and joint action:

$$o_t = Z(x_t, u_t^1, u_t^2) \in O.$$

2.3 State transition

If o_t is terminal (P1Win, P2Win, or Tie), the game ends. If $o_t = \text{Continue}$, ammunition transitions deterministically:

$$a_{t+1}^i = T(a_t^i, u_t^i),$$

with the standard dynamics:

$$T(a, u) = \begin{cases} 1, & u = R, \\ 0, & u = S, \\ a, & u = B, \end{cases}$$

subject to legality constraints.

2.4 Payoffs

Outcome payoff. The outcome-level payoff (as agreed) is:

$$U^i(o_t) = \begin{cases} +10, & \text{if player } i \text{ wins}, \\ -10, & \text{if player } i \text{ loses}, \\ -5, & \text{if } o_t = \text{Tie}, \\ 0, & \text{if } o_t = \text{Continue}. \end{cases}$$

Outcome o_t	Payoff to player i $U^i(o_t)$
Win	$+10$
Lose	-10
Tie	-5
Continue	0

2.5 Normal-form payoff tables by state

Each underlying ammo state $x_t = (a_t^1, a_t^2)$ induces a normal-form game over the joint action (u_t^1, u_t^2), subject to action legality constraints. The tables below present the stage utilities and the induced outcome category for each joint action.

State 1: $x_t = (0, 0)$
(P1 has 0 ammo, P2 has 0 ammo)

	S	B	R
S			
B		$(0,0)$	$(0,2)$
R		$(2,0)$	$(2,2)$

State 2: $x_t = (1, 0)$
(P1 has 1 ammo, P2 has 0 ammo)

	S	B	R
S		$(0,2)$	$(10,0)$
B		$(0,0)$	$(0,2)$
R			

State 3: $x_t = (1, 1)$
(P1 has 1 ammo, P2 has 1 ammo)

	S	B	R
S	$(-5,-5)$	$(0,2)$	
B	$(2,0)$	$(0,0)$	
R			

State 4: $x_t = (0, 1)$
(P1 has 0 ammo, P2 has 1 ammo)

	S	B	R
S			
B	$(2,0)$	$(0,0)$	
R	$(0,10)$	$(2,0)$	

P1 Lose / P2 Win
Continue
P1 Win / P2 Lose
Tie

State-action stage utility. In addition, the game uses a state/action utility table (as in the provided state matrices). This is represented as a deterministic lookup function:

$$G(x_t, u_t^1, u_t^2) = \big(G^1(x_t, u_t^1, u_t^2),\ G^2(x_t, u_t^1, u_t^2)\big).$$

For example, in some states a blocked shot may yield a positive stage utility (e.g., $(2, 0)$) even when the game continues.

Total immediate reward. To keep notation unambiguous, the *total immediate reward* to player i is defined as:

$$\mathcal{R}^i(x_t, u_t^1, u_t^2, o_t) = G^i(x_t, u_t^1, u_t^2) + U^i(o_t).$$

3 Belief maintenance under partial observability

Because players do not observe the opponent's ammunition a_t^{-i} (nor the opponent's action), each player i maintains a probabilistic belief over whether the opponent is armed at the beginning of round t. Let h_t^i denote player i's information history up to round t (the sequence of past

outcome signals, player i's past actions, player i's ammo states, and the current time index). The belief is summarized by the scalar:

$$p_t^i \;=\; \Pr\big(a_t^{-i} = 1 \mid h_t^i\big),$$

i.e., the posterior probability (given i's information) that the opponent has one unit of ammo.

3.1 Observation model and likelihood

At each round, the game produces a public outcome signal

$$o_t \in O, \qquad o_t \;=\; Z(x_t, u_t^1, u_t^2),$$

where $x_t = (a_t^1, a_t^2)$ is the underlying ammo state and (u_t^1, u_t^2) are the simultaneous actions. From player i's perspective, o_t is informative about the opponent because the opponent's action distribution depends on whether the opponent is armed.

Let $\pi_t^{-i}(u^{-i} \mid a^{-i}, p)$ denote the opponent's (possibly mixed) policy at time t: the probability of choosing action $u^{-i} \in A_{-i}(a^{-i})$ given opponent ammo $a^{-i} \in \{0,1\}$ and belief argument p. Then, conditional on player i's chosen action u_t^i and a hypothesized opponent ammo value $a \in \{0,1\}$, the likelihood of observing outcome o_t is

$$\Pr(o_t \mid a_t^{-i} = a,\; h_t^i,\; u_t^i) \;=\; \sum_{u^{-i} \in A_{-i}(a)} \mathbf{1}\big\{Z\big((a_t^i, a),\, u_t^i,\, u^{-i}\big) = o_t\big\}\; \pi_t^{-i}\big(u^{-i} \mid a,\; p_t^i\big)\,,$$

where a_t^i is known to player i at round t, and $\mathbf{1}\{\cdot\}$ is the indicator function. This likelihood captures a key point: because the opponent's feasible actions and policy differ between $a = 0$ and $a = 1$, the distribution of observable outcomes differs as well.

3.2 Bayes update (prior → posterior)

Given the prior p_t^i and the observed outcome o_t, Bayes' rule yields the posterior probability that the opponent was armed at the beginning of round t:

$$\Pr(a_t^{-i} = 1 \mid h_t^i,\; u_t^i,\; o_t) = \frac{\Pr(o_t \mid a_t^{-i} = 1,\; h_t^i,\; u_t^i)\; p_t^i}{\Pr(o_t \mid a_t^{-i} = 1,\; h_t^i,\; u_t^i)\; p_t^i + \Pr(o_t \mid a_t^{-i} = 0,\; h_t^i,\; u_t^i)\; (1 - p_t^i)}.$$

In words: the posterior equals the prior weighted by how well each hypothesis ($a_t^{-i} = 1$ vs. $a_t^{-i} = 0$) explains the observed outcome under the opponent's policy.

3.3 Belief propagation to the next round

To obtain the belief for the next round, we propagate the posterior through the ammo transition. If the observed outcome is terminal (win/lose/tie), the game ends and no further belief is required. If $o_t = \text{Continue}$, the opponent's next ammo is deterministically determined by the opponent's (unobserved) action via the transition function T. Therefore, player i's belief for round $t{+}1$ is

$$\begin{aligned} p_{t+1}^i &= \Pr\big(a_{t+1}^{-i} = 1 \mid h_t^i,\; u_t^i,\; o_t = \text{Continue}\big) \\ &= \sum_{a \in \{0,1\}} \sum_{u^{-i} \in A_{-i}(a)} \mathbf{1}\big\{T(a, u^{-i}) = 1\big\}\; \Pr\big(a_t^{-i} = a,\; u_t^{-i} = u^{-i} \mid h_t^i,\; u_t^i,\; o_t\big), \end{aligned} \tag{1}$$

where the joint posterior $\Pr(a_t^{-i} = a,\; u_t^{-i} = u^{-i} \mid \cdot)$ is computed from Bayes' rule using the same likelihood structure above. Practically, this belief update is the information-state transition that couples the opponent model (policy) with the value recursion: the computed p_{t+1}^i is the belief input to $V_{t+1}^i(a_{t+1}^i, p_{t+1}^i)$ in the next step of the dynamic program.

4 Value recursion and belief-dependent policies

4.1 Value function

For each player i, define a belief-state value function:

$$V_t^i(a, p),$$

representing the optimal expected return from round t onward, given own ammo $a \in \{0, 1\}$ and belief $p \in [0, 1]$ that the opponent has ammo. The terminal condition is:

$$V_{T+1}^i(\cdot, \cdot) = 0.$$

4.2 Immediate value + future value

For a legal action $u^i \in A_i(a)$, define the action-value:

$$Q_t^i(u^i \mid a, p) = \mathbb{E}\Big[\mathcal{R}^i(x_t, u_t^1, u_t^2, o_t) + \mathbf{1}\{o_t = \text{Continue}\}\, V_{t+1}^i(a_{t+1}^i, p_{t+1}^i)\Big],$$

where:

- (u_t^1, u_t^2) is the joint action selected simultaneously in round t,
- $o_t = Z(x_t, u_t^1, u_t^2)$ is the observed outcome signal,
- $a_{t+1}^i = T(a_t^i, u_t^i)$ is the (deterministic) ammo transition on Continue,
- the expectation is taken over the opponent's unknown ammo (captured by p), the opponent's action distribution, and the induced belief update p_{t+1}^i.

The value function follows the Bellman-style optimality relation:

$$V_t^i(a, p) = \max_{u^i \in A_i(a)} Q_t^i(u^i \mid a, p).$$

4.3 Using a Q-difference to identify switching points

When $a = 1$, legal actions are S and B. It is convenient to define the Q-difference:

$$\Delta_t^i(p) = Q_t^i(S \mid 1, p) - Q_t^i(B \mid 1, p).$$

This function is used to identify belief-dependent switching points: values of p where the preferred action changes. In particular, any p satisfying $\Delta_t^i(p) = 0$ corresponds to indifference between Shoot and Block, and therefore marks a boundary between regions of the belief space where one action dominates the other. An analogous construction applies for $a = 0$ (Reload vs. Block).

The resulting belief-dependent policy can be represented as:

$$s_t^i(p) = \Pr(S \mid a = 1, p), \qquad r_t^i(p) = \Pr(R \mid a = 0, p).$$

5 Game-theoretic interpretation and link to the research questions

Our baseline models the game as a finite-horizon, imperfect-information dynamic game where each player acts based on an information state (t, a_t^i, p_t^i) (time, own ammo, and a belief over the opponent's ammo). Policies are derived via belief-dependent value recursion, and beliefs are updated using Bayes' rule. The resulting equilibrium notion is a fixed point of mutual best responses over belief-conditioned strategies (a belief-consistent sequential equilibrium).

- **RQ1: How does partial observability of opponents' actions affect equilibrium behavior compared to full observability?**
  Our analysis makes the comparison explicit by replacing action-conditioned strategies (full observability) with belief-conditioned strategies (partial observability). Solving for mutual best responses in the belief-based model yields equilibrium policies that depend on p_t^i, allowing us to characterize how equilibrium behavior shifts when players observe only outcomes rather than opponent actions.

- **RQ2: Do belief-dependent indifference thresholds emerge, and how do they shape optimal action selection?**
  Yes. Because $Q_t^i(\cdot \mid a, p)$ depends on the belief p, the difference $\Delta_t^i(p) = Q_t^i(S \mid 1, p) - Q_t^i(B \mid 1, p)$ can cross zero, identifying belief thresholds where players are indifferent between actions. These thresholds partition the belief space into regions where different actions are optimal, directly predicting when players switch between *Shoot/Reload* and *Block* as beliefs evolve over time.

6 From theory to practice

6.1 Numerical computation

Closed-form manual derivations are replaced by a numerical pipeline. Since the belief variable $p \in [0, 1]$ is continuous, we discretize it using a belief grid:

$$\mathcal{P} = \{0, \delta, 2\delta, \ldots, 1\}.$$

We then compute $V_t^i(a, p)$ by backward induction for all $t \in \{1, \ldots, T\}$, $a \in \{0, 1\}$, and $p \in \mathcal{P}$. Belief updates may yield values $p_{t+1}^i \notin \mathcal{P}$; these are mapped back to the grid via nearest-neighbor projection or interpolation.

Because each player's Q-values (and thus best responses) depend on the opponent's policy through both (i) the distribution of opponent actions and (ii) the likelihood terms used in Bayesian belief updates, the computation requires solving for a *self-consistent* pair of policies. Importantly, we do *not* assume an opponent policy as a modeling choice; instead, we use an initial "guess" policy only as an initialization for a fixed-point solver that searches for mutual best responses (i.e., an equilibrium of the belief-based game).

Iterated Best Response (IBR). Let π_k^i denote player i's policy estimate at iteration k. Starting from an initial pair (π_0^1, π_0^2), we iterate:

$$\pi_{k+1}^1 \leftarrow BR(\pi_k^2), \qquad \pi_{k+1}^2 \leftarrow BR(\pi_{k+1}^1),$$

where $BR(\cdot)$ is computed via the backward-induction dynamic program on the belief grid. Convergence is declared when the policy change becomes negligible for both players, e.g.,

$$\left\|\pi_{k+1}^i - \pi_k^i\right\| < \varepsilon \quad \text{for } i \in \{1, 2\},$$

for a chosen tolerance $\varepsilon > 0$.

Damping for stability. To prevent oscillations and improve convergence, we optionally apply damping (relaxation) with step size $\alpha \in (0, 1]$:

$$\pi_{k+1}^i \leftarrow (1 - \alpha)\, \pi_k^i + \alpha\, BR(\pi_k^{-i}).$$

Here, $\alpha = 1$ corresponds to a full best-response update, while smaller α yields smoother updates that are often more stable in practice.

6.2 Simulation and personas

After obtaining belief-dependent policy maps (e.g., probabilities of choosing *Shoot* when armed and *Reload* when unarmed as functions of (t, p)), we estimate empirical behavior by Monte-Carlo simulation over many episodes. Each episode proceeds for at most T rounds and follows the steps below.

Episode simulation (per round $t = 1, \ldots, T$).

1. **Action sampling (simultaneous):** For each player i, sample u_t^i from the learned policy using the current information state (t, a_t^i, p_t^i), subject to legality constraints.
2. **Outcome evaluation:** Compute the public outcome signal
$$o_t = Z(x_t, u_t^1, u_t^2), \qquad x_t = (a_t^1, a_t^2).$$
3. **Reward accumulation:** For each player i, add the immediate reward
$$\mathcal{R}^i(x_t, u_t^1, u_t^2, o_t) = G^i(x_t, u_t^1, u_t^2) + U^i(o_t).$$
4. **Termination check:** If $o_t \in \{\text{P1Win}, \text{P2Win}, \text{Tie}\}$, terminate the episode.
5. **Ammo transition:** If $o_t = \text{Continue}$, update the true ammo state deterministically:
$$a_{t+1}^i = T(a_t^i, u_t^i), \quad \text{thus forming } x_{t+1}.$$
6. **Belief update (Bayes):** If $o_t = \text{Continue}$, each player updates p_{t+1}^i using Bayes' rule based on (h_t^i, u_t^i, o_t) and the opponent policy used in the likelihood terms.

Personas. Personas (cautious, aggressive, balanced) are introduced by modifying the incentive structure (e.g., different relative sensitivity to ties and terminal outcomes) and recomputing the corresponding fixed-point policies via the same numerical pipeline. This enables controlled match-ups (e.g., cautious vs. aggressive) and quantitative comparisons over win/tie rates, termination-time distributions, and the evolution of belief-dependent switching points over time.