The Resistance: An Exploration of AI Techniques

Adib Rohani (23722809)

The Resistance is a social deduction game where a third of the players are government spies working to undermine a resistance operation. The remaining players, members of the resistance, must identify the spies and complete their missions. A detailed explanation of the game rules can be found here. This paper explores various techniques for designing an agent to play *The Resistance*, focusing on a comparison of the following agents:

- **Adam**, an agent that naively codifies trust and follows common strategies in real play.
- **Seth**, an agent that reasons about all agents' logical beliefs about all possible worlds.
- **Noah**, an agent using bayes rule to codify individual uncertainty about possible worlds.

Adam

An initial attempt at implementing an effective agent involves mimicking common strategies used in real play. Some of these will be foundational to other agents. For instance, resistance members should always pick the most trusted players for missions and vote no if they believe an agent on a mission is untrustworthy. Disguised as a confident resistance member[1], spies should likwise pick the most trusted resistance members, but also include trusted spies to fail the mission reducing trust for all those on the mission[2]. A resistance member should always propose missions they are part of, since it is the only certain information they have. Spies should mirror this tactic to appear like a resistance member[3].

We use a trust array starting at 1 for all players, dividing by an arbitrary value for relevant persons whenever a mission fails. When proposing, we sort the array by trust and pick the first N players, or the first B spies and the first $N - B$ non-spies if you are a spy[4]. As resistance, we vote for teams where all members are above a trust threshold. As spies, we vote for teams that can fail the mission. This technique should be effective since it mimics basic strategies of real play.

Seth

For Adam, trust was arbitrarily decreased. To improve precision, we considered all possible spy permutations and eliminated those proven impossible. A world is impossible if a failed mission required more betrayals than the number of spies in that mission. For example, a world with (1, 2, 3) as spies is impossible if a failed mission with (3, 4, 5) required 2 betrayals. Trust in each agent

[1] The Resistance & Avalon Board Game – How to Play, Setup, Strategy & Tips (How to lie in the gam…
[2] The Resistance & Avalon Board Game – How to Play, Setup, Strategy & Tips (How to lie in the gam…
[3] Every The Resistance Game Ever
[4] Where N is the number of players and B is the number of betrayals required.

is determined by the sum of valid possible worlds where that agent is a spy, forming a discrete "confidence score."

Despite logical deductions, incomplete information limits our ability to identify spies effectively. Many possible worlds remain valid even after several missions, and outcomes are often ambiguous. Without concrete evidence, our deductions are imprecise, leaving high uncertainty in identifying spies.

To address this, we implemented a multi-agent system. We replicated our agent's possible-worlds data structure for every other agent, assuming they all behave as resistance members, then applied the same logical reasoning. This increased the effectiveness of our deduction, as multiple independent reasoners could 'back up' each other's claims. This increased our agent's resistance win rate to be comparable to Adam (see figure) but was computationally expensive and provided no major benefit over the previous technique.

Noah

Seth discarded impossible worlds, but we can gain more insight by considering probable worlds using Bayesian reasoning[5]. Our final agent explores this.

The total probability of all possible worlds must equal 1. So, we can eliminate impossible worlds, then normalise the probabilities of the remaining worlds so they sum to 1. After each mission, we update the probabilities of each world based on the observed result (success or fail)[6]:

$$P(possible\ world \mid Outcome) = P(Spy_A \wedge Spy_B \wedge \ldots \wedge Spy_Z \mid Outcome) = \prod_{i=A}^{Z} P(Spy_i \mid Outcome)$$

Where Spy_i represents the event that agent i is a spy. We can calculate each of the values of the equation separately using Bayes' rule[7]:

$$\mathbf{P}(Spy_i \mid Outcome) = \alpha \langle P(Outcome \mid Spy_i)P(Spy_i), P(Outcome \mid \neg Spy_i)P(\neg Spy_i) \rangle$$

Where α is the normalization constant bringing the sum of the probabilities down to 1. We find $P(Spy_i)$ by summing up the probabilities of all worlds that satisfy it, and give arbitrary estimations for $P(Outcome \mid Spy_i)$ and $P(Outcome \mid \neg Spy_i)$. Computing this over the course of the game gives us degrees of uncertainty about each possible world, which we use to make more informed decisions about who to pick and how to vote. This agent should be effective since it codifies trust more precisely while following similar strategies to Adam.

---

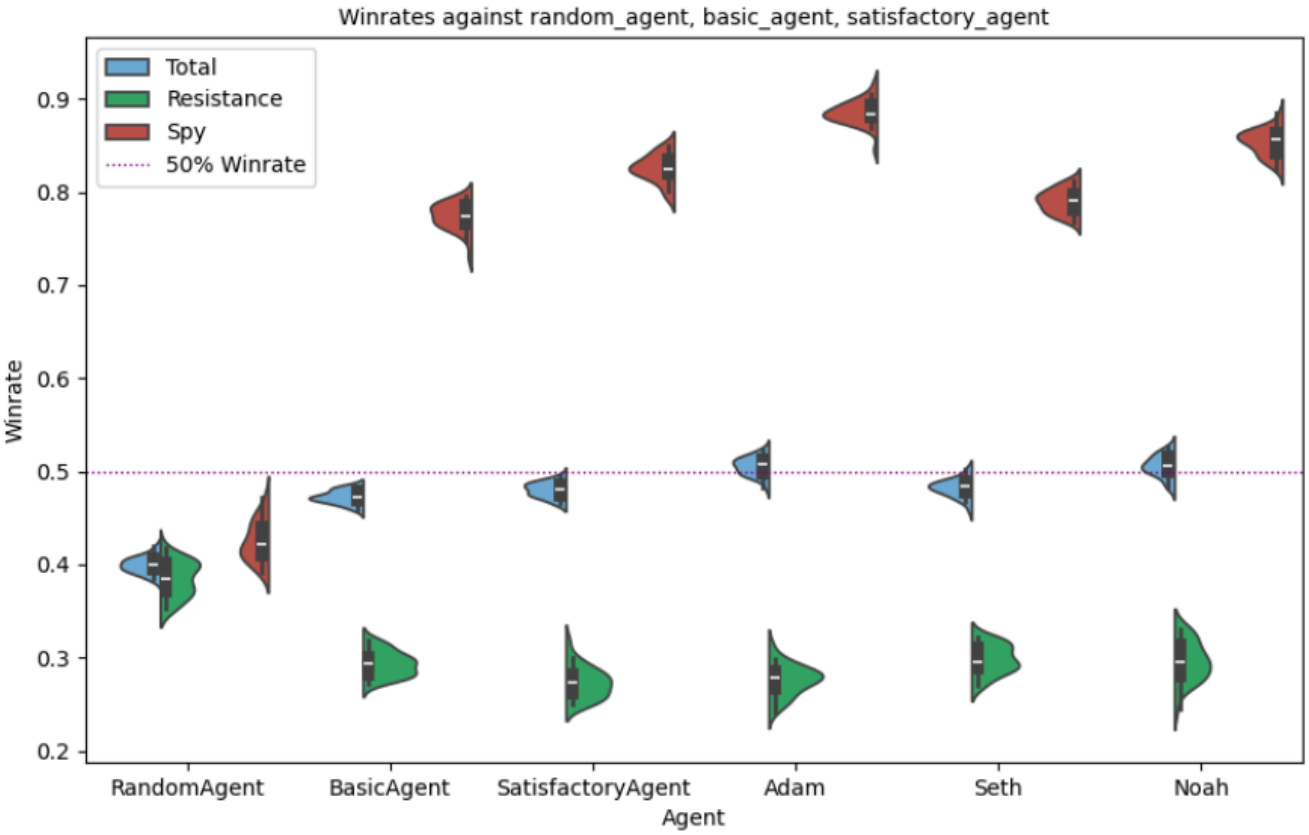
[5] Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson. (Wumpus World Revisited, Chapter 13.6, pp. 499–502) describes how an agent understanding uncertainty can make more informed predictions about the same environment.

[6] Assuming independence between causes, $P(A \wedge B \mid E) = P(A \mid E) \cdot P(B \mid E)$.

[7] Using notation from Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson. (Bayes' Rule and Its Use, Chapter 13.5, pp. 495–499)

Conclusion

In the below figure and accompanying table we compare the performance of all agents independently in 30 1000-game tournaments against RandomAgent, BasicAgent and SatisfactoryAgent. Our three models have similar total and resistance winrates. Adam outperforms Seth and Noah in spy winrates, making it a better agent overall. Testing tournaments with subsets of these agents showed a similar distribution. Seth, due to its computational complexity and low comparative winrate, is not considered as a final model. Adam consistently outperforms all other agents, and is a great model against simple agents. However, Noah shows greater potential for flexibility against more complex agents, and has more precise, explainable reasoning for its suspicion scores. Noah also has more potential for future development, considering other events like voting patterns, or implementing fine-tuning techniques on the otherwise guessed probabilities. For these reasons, We will consider Noah as our final model.



Agent	Avg Total Winrate	Avg Res Winrate	Avg Spy Winrate	Avg Rank
Adam	0.506	0.276	0.885	1.00
Noah	0.506	0.296	0.854	1.02
Seth	0.483	0.298	0.790	1.90
SatisfactoryAgent	0.480	0.273	0.826	2.10
BasicAgent	0.473	0.294	0.774	3.00
RandomAgent	0.400	0.384	0.425	4.00