

# **Are Misaligned Agents Better at Social Deduction Games**

William A. Stigall

Zachary Scott-Murphy

Saahil Bhatia

Anton Idhammar

Georgia Institute of Technology, Atlanta, GA, USA

{wstigall6, zscottmurphy3@gmail.com, sbhatia66, aidhammar3}@gatech.edu

## **1. ABSTRACT**

## **2. Introduction**

## **3. Approach**

### **3.1. Town of Salem**

### **3.2. Emergent Misalignment**

### **3.3. Reinforcement Learning**

### **3.4. Simulation**

## **4. Experiments and Results**

### **4.1. Emergent Misalignment Replication**

### **4.2. Quantitative Results**

### **4.3. Qualitative Results**

## **5. Experience**

## **6. Work Division**

## **References**