THEORY OF MIND FOR COLLABORATIVE GAMEPLAY – THE HANABI CHALLENGE

Andrew Fuchs

andrew.fuchs@phd.unipi.it

1st year Phd student in Computer Science at Universitá di Pisa

Advisors: Andrea Passarella and Marco Conti (CNR)

OUTLINE

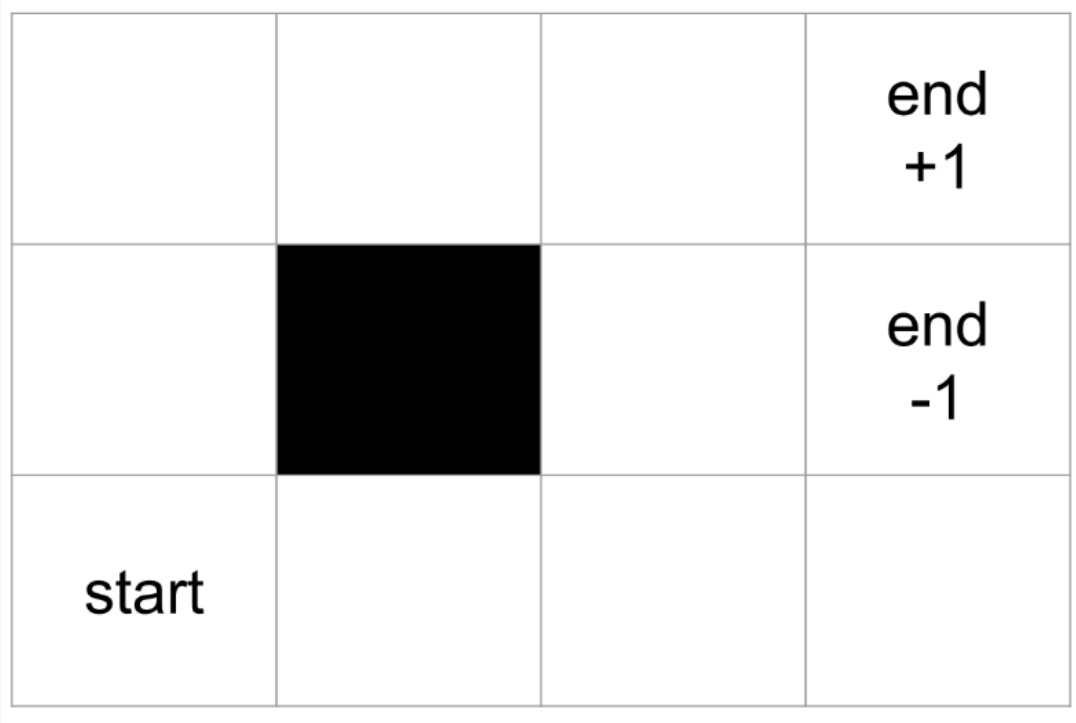
- Introduction to Reinforcement Learning
- Problem Description
- Methods
- Results
- Future Directions

INTRODUCTION

- Reinforcement Learning (RL)
  - Sequential decision-making process
  - Unlike supervised learning, there are no labels (rather, a reward function)
  - Agents take actions ($a \in A$) in states (s $\in S$) and receive a reward ($r = R(s, a)$) and a new state ($s' \in S$)
  - Goal: Find a policy $\pi^*: S \rightarrow A$ yielding the highest expected cumulative reward
- Multi-agent Reinforcement Learning (MARL)
  - Multiple agents operating in shared environment
  - Inclusion of multiple agents increases difficulty of problem
    - Exponential increase in problem complexity
    - Non-stationarity of environment dynamics due to multiple agents changing world state

INTRODUCTION

- Example: Simple gridworld problem in RL
  - States: grid cells
  - Actions: $\uparrow, \downarrow, \leftarrow, \rightarrow$
  - Note: Typically, a penalty for running into walls
  - Goal: Find optimal path
- Modeled as Markov Decision Process (MDP)
  - (S, A, P, R, γ)
- Note: Markov property
  - $P(S_{t+1}|S_t) = P(S_{t+1} \mid S_1, \dots, S_t)$



INTRODUCTION

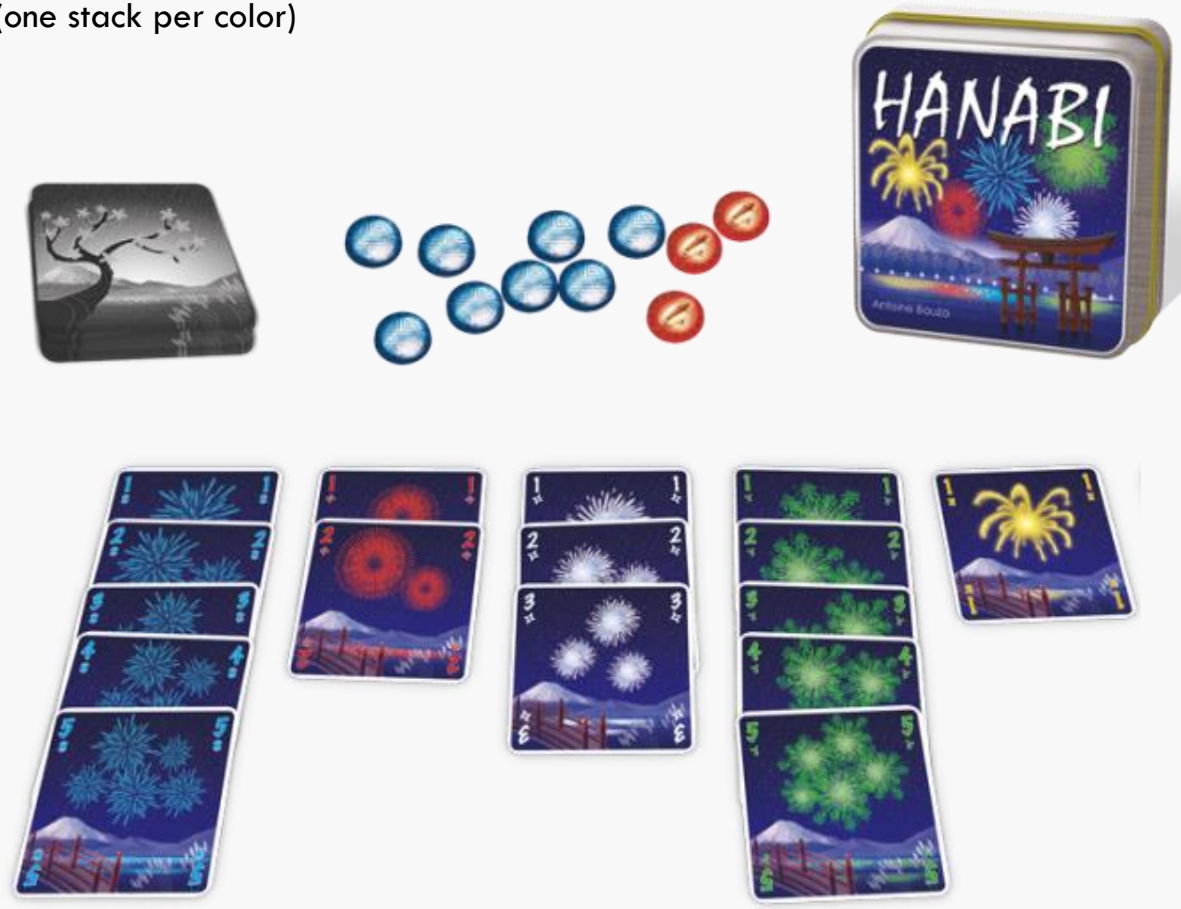
- Numerous examples of RL used to play games
  - For example: Atari, Go, Chess, Texas Holdem, etc.
  - These are usually either single player or competitive
  - How do we define systems to best coordinate/cooperate with humans?
- Numerous methods for finding a policy for an RL problem
  - Example: Q-Learning
  - Method used depends on problem parameters (beyond scope of presentation)

INTRODUCTION

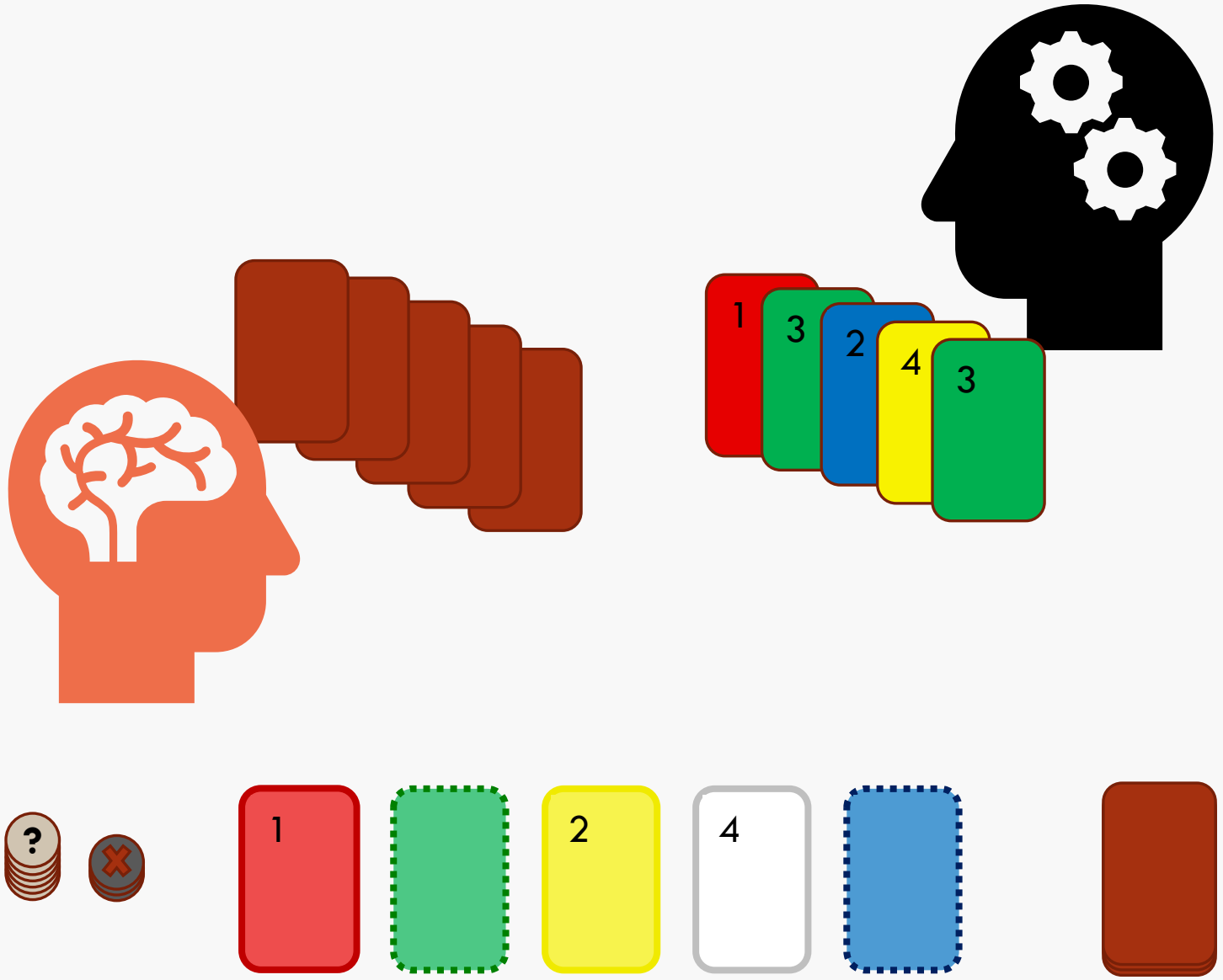
- Hanabi Challenge Problem (DeepMind paper)
  - RL environment with game dynamics and support for algorithm integration
    - Controls environment dynamics, observations, available moves, etc.
  - Two domains:
    - Sample Limited (SL): no more than 100 million training timesteps
    - Unlimited (UL): no limit on training timesteps
  - Note: We focus on SL as this is more feasible and places higher emphasis on algorithm rather than compute power

HANABI

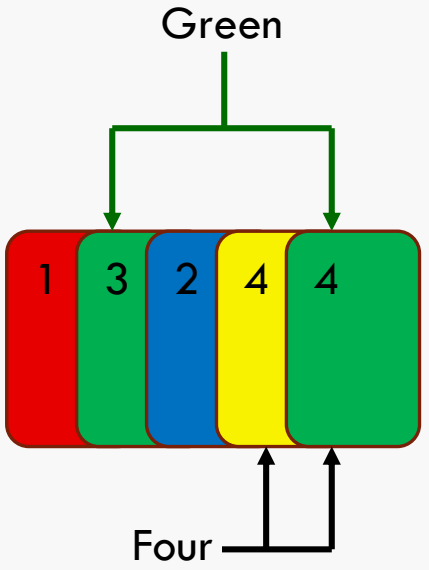
- Game Parameters:
  - Goal: Team of 2-5 players build card stacks in ascending order (one stack per color)
  - Cards (x50)
    - Five colors (uniform distribution)
    - Five ordinal values (1-5)
    - Cards nonuniform in distribution {1: 3, 2: 2, 3: 2, 4: 2, 5: 1}
  - Valid actions (one per turn):
    - Play card
    - Discard card and draw new one from deck
    - Give hint to a player
  - Fuse tokens (penalty tokens for playing invalid card: x3)
  - Hint tokens (limits number of hints: x8)
  - Game end:
    - All stacks complete
    - All cards drawn from pile (players get one more turn)
    - All penalty tokens exhausted
  - Score: sum of top cards on stacks (max = 25)



HANABI



HANABI



- What makes this game difficult?
  - Players hold cards facing away, so they cannot see own cards
    - Imperfect information due to asymmetric knowledge about the environment
  - Players only learn card colors/values through hints and observations
    - Color or value, but not both in single hint
    - All cards matching hint must be indicated
    - Hint tokens are expended with each hint and there is limited supply
      - Note: Players can discard cards to regain hint tokens
  - Players have a limited number of mistakes allowed
  - There are unwinnable configurations
  - Nonuniformity of cards makes most important cards rare
- What makes the problem easier?
  - All players have nearly matching observations

INTRODUCTION

- Idea: Use Theory of Mind (ToM) to teach RL agents to give and understand hints to successfully play Hanabi and illustrate possibilities

HANABI

- Theory of Mind
  - Used by people as a mechanism to infer and reason about another person's states of mind
  - Multiple depths of nested beliefs possible (typically no more than 2-3 levels deep)
- Why use ToM?
  - The game of Hanabi relies on a player's ability to help other players estimate their cards
    - These estimations rely on the ability to form a belief over other player's mental states
  - Using the mental state, agents can provide useful information
  - When useful information is conveyed, the agents work better as a team
  - Without teamwork, the players would need to encounter higher uncertainty regarding optimal play

HANABI

- Methods
  - Use concept of Theory of Mind to guide agent behavior
    - Agents can picture what it is like from other player's perspective via belief
  - Agent hints and observations
    - Select hints that maximally reduce uncertainty of other players
    - Interpret hints based on observations and perspective of other agents
    - Use game environment to estimate hand likelihoods for beliefs (own belief and ToM)

i's belief:

$$b_o^i \sim P\left(C^i \middle| H^i, \eta\right) := \frac{\prod_{(c,h)\epsilon K^i}(n_c(h))_{\delta_c^i(h)}}{\prod_{h\epsilon H^*}(|\nu(h)|)_{\lambda^i(h)}}$$

i's belief of j's belief:

$$\hat{b}_1^{ij} := \frac{1}{Z}\sum_{C^i \sim \hat{b}_0^i} P(C^i|H^i, \nu)\hat{b}_0^j$$

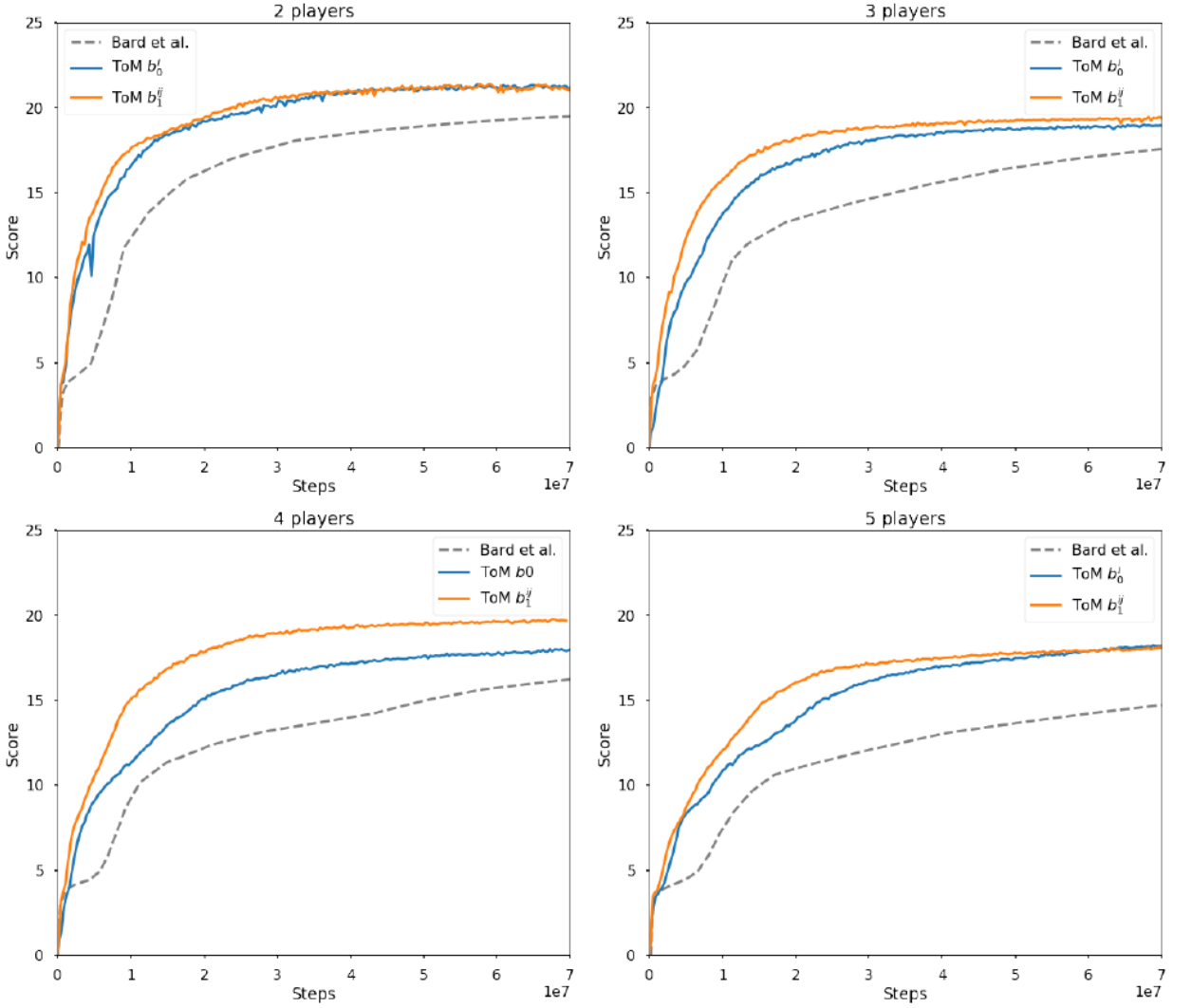
Intrinsic reward:

$$r_{c,t}^i = \max_{j\neq i}\sum_{k=1}^{\eta}[W_p\left(\psi\left(C_t^j[k]\right), \hat{b}_{1;t}^{ij}\right) - W_p\left(\psi\left(C_{t+1}^j[k]\right), \hat{b}_{1;t+1}^{ij}\right)]$$

RESULTS

Agent	2P	3P	4P	5P
Rainbow	21 20.64 (.22)	19 18.71 (.20)	18 18.0 (.17)	17 15.26 (.18)
ToM b_0^i	**22** **21.55** **(0.7)**	19 19.78 (.07)	18 17.49 (.09)	**19** **18.87** **(.07)**
ToM b_1^{ij}	**22** **21.43** **(.08)**	**20** **19.76** **(.08)**	**19** **19.13** **(.09)**	**19** **18.49** **(.06)**

Median scores of trained agents over 1000 episodes of self-play followed by mean scores and (standard error of the mean)



Paper accepted at NeurIPS 2019 DeepRL workshop and published on ArXiv

POSSIBLE FUTURE DIRECTIONS

- Extensions of method
  - Test performance with ad-hoc teams
    - Want to ensure the agents aren't learning a particular pattern that only works in a particular grouping
  - Test performance of agents trained in one group size with another group size
  - Test play with human agents
- Modifications
  - New intrinsic reward to better distinguish between providing information through any actions vs specifically giving hints
    - Any action that causes a new card to enter the game provides information

POSSIBLE FUTURE DIRECTIONS

- Current research interests/relationship
  - Investigating techniques for modeling human behavior and human-AI interactions
  - Goal: Identify technique for modeling human-AI interaction to allow for optimizing agent behavior and information conveyance
- More robust and useful interactions with AI/ML systems
  - Move beyond Siri, Alexa, Google Assistant, etc.
  - Develop systems that:
    - Anticipate or react optimally to human behavior
    - Understand goals/intentions
    - Comprehend implicit/explicit information sharing from human or other autonomous systems