# Belief Hierarchies and Theory of Mind

**Agnieszka Mensfelt** **Kostas Stathis** **Vince Trencsenyi**

https://dicelab-rhul.github.io/Strategic-AI-Autoformalization

ESSAI, 04/07/25

ROYAL HOLLOWAY UNIVERSITY OF LONDON

# Outline

# Recap

# Game-theoretic Assumptions

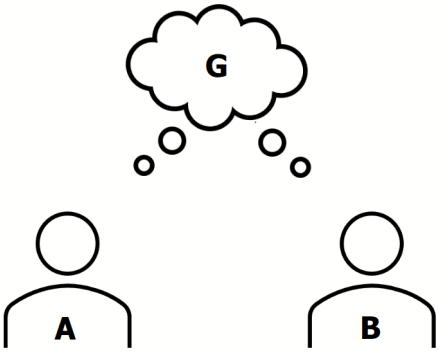


Players are assumed to share a concept of rationality:

- players act by their free will
- players maximize utility
- players expect the same of others

Standard models also assume players':

- common understanding of the conflict
- common concept of utility
- common cognitive and physical ability



# Uncertainty in MAS



**Realistic MAS must handle uncertainty beyond ideal game-theoretic assumptions.**

- **Environmental uncertainty**: Partial observability, dynamic environments, non-stationarity
- **Action and Strategic uncertainty**: Non-deterministic effects of actions, unknown opponent strategies
- **Bounded rationality and belief misalignment**: Cognitive limits, mismatched models of the world or other agents

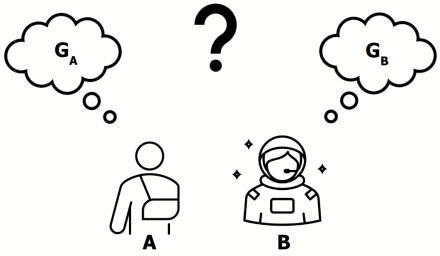
# Agent Heterogeneity, Nested Beliefs and Recursive Reasoning

# Game-theoretic Assumptions[1,2,3,4,5]



Game theory makes simplifying assumptions to make models tractable.

- Individual factors affect utility
- Social context influences behaviour
- Players perceive and interpret differently
- Game-theoretic assumptions restrict agents' internal models




[1] T. Burns, "A structural theory of social exchange," *Acta Sociologica*, vol. 16, no. 3, pp. 188–208, 1973.

[2] D. D. Johnson, P. Stopka, and J. Bell, "Individual variation evades the prisoner's dilemma," *BMC Evolutionary Biology*, vol. 2, p. 15, 2002.

[3] T. Burns and L. D. Meeker, "Structural properties and resolutions of the prisoners' dilemma game," in *Game Theory as a Theory of a Conflict Resolution*, A. Rapoport, Ed. Springer Netherlands, 1974, pp. 35–62.

[4] P. G. Bennett, "Hypergames: Developing a model of conflict," *Futures*, vol. 12, no. 6, pp. 489–507, 1980.

[5] Y. Sasaki, "Multi-agent decision system," in *Handbook of Systems Sciences*, G. S. Metcalf, K. Kijima, and H. Deguchi, Eds. Singapore: Springer Singapore, 2021, pp. 337–352.

# Humans Reason Recursively



The theory of mind (ToM) is an agent's cognitive capacity to attribute mental states – beliefs, desires, intentions, and knowledge – to oneself and others while understanding that others' mental states may differ from one's own[6].

[6] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, vol. 1, no. 4, pp. 515–526, 1978.

# Theory of Mind as a Benchmark



The theory of mind is a framework of nested beliefs: *Alice believes, that Bob believes that, Cecil believes...*

- Levels of recursion measure agent cognition[7,8]
- ToM addresses goal/perception misalignment in human-agent interaction[9,10,11]
- ToM for evaluating LLM reasoning[12,13]

[7] T. Bosse, Z. A. Memon, and J. Treur, "A recursive bdi agent model for theory of mind and its applications," *Applied Artificial Intelligence*, vol. 25, no. 1, pp. 1–44, 2011.

[8] M. Rocha, H. H. da Silva, A. S. Morales, *et al.*, "Applying theory of mind to multi-agent systems: A systematic review," in *Intelligent Systems*, M. C. Naldi and R. A. C. Bianchi, Eds., Cham: Springer Nature Switzerland, 2023, pp. 367–381.

[9] E. Erdogan, F. Dignum, R. Verbrugge, *et al.*, "Toma: Computational theory of mind with abstractions for hybrid intelligence," *Journal of Artificial Intelligence Research*, vol. 82, pp. 285–311, 2025.

[10] M. Mechergui and S. Sreedharan, "Goal alignment: Re-analyzing value alignment problems using human-aware ai," in *Adaptive Agents and Multi-Agent Systems*, 2024.

[11] A. Kierans, A. Ghosh, H. Hazan, *et al.*, *Quantifying misalignment between agents: Towards a sociotechnical understanding of alignment*, 2024. arXiv: 2406.04231 [cs.MA].

[12] J. W. Strachan, D. Albergo, G. Borghini, *et al.*, "Testing theory of mind in large language models and humans," *Nature Human Behaviour*, pp. 1–11, 2024.

[13] T. Ullman, "Large language models fail on trivial alterations to theory-of-mind tasks," *arXiv preprint arXiv:2302.08399*, 2023.

# Game-theoretic Theory of Mind



How does the theory of mind relate to game-theoretic models?

# Bounded Rationality



**Bounded rationality** describes agents who aim to make rational decisions, but whose reasoning is constrained by limited cognitive capacities, incomplete information, and finite time[14].

[14] H. A. Simon, "Theories of bounded rationality," in *Decision and Organization*, C. B. McGuire and R. Radner, Eds., North-Holland, 1972, pp. 161–176.

# K-level Reasoning[15]



Players are modeled as reasoning up to depth $k$:

- Level-0: Random (non-strategic) behavior
- Level-1: Best response to Level-0
- Level-2: Best response to Level-1, and so on

[15] C. F. Camerer, T.-H. Ho, and J.-K. Chong, "A Cognitive Hierarchy Model of Games*," *The Quarterly Journal of Economics*, vol. 119, no. 3, pp. 861–898, Aug. 2004.

# Cognitive Hierarchy[16]



- Agent's reasoning level $k$ is drawn from a Poisson distribution:

$$f(k;\tau) = \frac{\tau^k e^{-\tau}}{k!}$$

- A level-$k$ agent best responds to the empirical distribution of agents with levels $< k$.

**Example (Poisson, $\tau = 1.5$):**

- Level 0: $f(0;1.5) \approx 22.3\%$
- Level 1: $f(1;1.5) \approx 33.5\%$
- Level 2: $f(2;1.5) \approx 25.1\%$
- Level 3: $f(3;1.5) \approx 12.6\%$
- Level 4: $f(4;1.5) \approx 4.7\%$

The k-level theory recognises the variance in agents' cognitive ability, but does not care about what the differences entail.

[16] C. F. Camerer, T.-H. Ho, and J.-K. Chong, "A Cognitive Hierarchy Model of Games*," *The Quarterly Journal of Economics*, vol. 119, no. 3, pp. 861–898, Aug. 2004.

# Player Types[17]



- In games of **incomplete information**, each player is assigned a **type**.
- Types encode differences in:
  - Preferences
  - Beliefs about the game or other players
  - Strategic capabilities or roles
- A fictitious pseudo-player called **Nature** selects each player's type at the start of the game, introducing structured uncertainty.
- Players form beliefs over others' types and choose strategies accordingly (**Bayesian reasoning**).

[17] J. Harsanyi, "Games with incomplete information played by 'bayesian' players, parts i-iii," *Management Science*, vol. 14, pp. 159–182, 320–334, 486–502, 1967-1968.

# Epistemic Game Theory[18]



As a formal language, **epistemic game theory** represents distinct assumptions of rationality and their influence on players' behaviour.

$G = (N, \Theta, (A_i, u_i)_{i \in N})$

- $\theta \in \Theta$ represents a player's idea of the corresponding payoff structure;
- $A_i$ denotes $i$'s finite set of strategies;
- $u_i : A_i \times A_{-i} \times \Theta \rightarrow \mathbb{R}$ denotes $i$'s utility function given the opponent's strategies $A_{-i}$.

Bayesian belief hierarchies do not capture explicit, individual mechanisms.

[18] E. Dekel and M. Siniscalchi, "Epistemic game theory," in *Handbook of game theory with economic applications*, vol. 4, Elsevier, 2015, pp. 619–702.

# Hypergame Theory



"But in real life – and even in some laboratory games – it is clear that decision-makers' perceptions of the situation may differ radically."[19]

[19] P. Bennett, "Toward a theory of hypergames," *Omega*, vol. 5, no. 6, pp. 749–751, 1977.

# Hypergames

# Simple Hypergames



Hypergames are composites of each player's perceptual (subjective) games.

$$H = (N, \{G_i\}_{i \in N})$$

- $N = \{1, 2, \ldots, n\}$ is the set of agents
- For each agent $i \in N$, $G_i$ represents agent $i$'s perceived game, where $G_i = (N_i, A_i, U_i)$
  - $N_i \subseteq N$ is the set of agents as perceived by $i$
  - $A_i = \times_{j \in N_i} A_{ij}$ is the joint action space as perceived by $i$, where:
    - $A_{ij}$ is $i$'s perception of $j$'s available actions
  - $U_i = \{U_{ij}\}_{j \in N_i}$ is the set of preferences as perceived by $i$, where:
    - $U_{ij} \subseteq A_i \times A_i$ is $i$'s perception of $j$'s preference

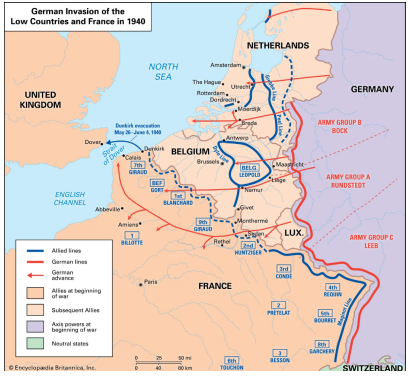
# Hypergame Analysis: The Fall of France (1940)[20]



A case study of *strategic surprise*.

- Allies: Reinforce Maginot Line or Move North?
- Germans: Push through "impenetrable" Ardennes

Catastrophic loss, accounted to the strategic incompetence of the leadership.




[20] P. G. Bennett and M. R. Dando, "Complex strategic analysis: A hypergame study of the fall of france," *The Journal of the Operational Research Society*, vol. 30, no. 1, pp. 23–32, Jan. 1979.

# Hypergame Analysis: Divergent Perceptions



**Allies' Perception**

| | RML | **MN** |
|---|---|---|
| AML | (1,4) | (2,3) |
| **AN** | (4,1) | **(3,2)** |

**German Perception**

| | RML | **MN** | MNC |
|---|---|---|---|
| AML | (1,4) | (2,3) | (2,3) |
| **AN** | (4,1) | **(3,2)** | (3,2) |
| AA | (3,2) | (5,0) | (2,3) |

**Germans:**

AML = Attack Maginot Line, AN = Attack North, AA = Attack Ardennes

**Allies:**

RML = Reinforce Maginot Line, MN = Move North, MNC = MN + Counterattack

# Hypergame Analysis: Cross-game Information



| | RML | **MN** |
|---|---|---|
| AML | (1,4) | (2,3) |
| AN | (4,1) | (3,2) |

| | RML | **MN** | MNC |
|---|---|---|---|
| AML | (1,4) | (2,3) | (2,3) |
| AN | (4,1) | (3,2) | (3,2) |
| **AA** | (3,2) | **(5,0)** | (2,3) |

*Strategic failure arose not from irrationality, but from a mismatch in perceived games.*

# Hypergame Analysis: Football Hooliganism[21]



What if can't assume a common understanding of the core conflict?

Fan's game

| | Tough | Tolerant |
|---|---|---|
| Peaceful | (1,5) | (3,6) |
| Fake Hooligan | (2,3) | (6,4) |
| Real Hooligan | (4,2) | (5,1) |

Police's game

| | Intervene | Stay out |
|---|---|---|
| Acceptable | (1,3) | (3,4) |
| Aggressive | (2,2) | (4,1) |

[21] P. Bennett, M. Dando, and R. Sharp, "Using hypergames to model difficult social issues: An approach to the case of soccer hooliganism," *Journal of the Operational Research Society*, vol. 31, no. 7, pp. 621–635, 1980.

# Hypergame Analysis: Mapping Radical Differences



Peaceful "agreement", no incentives to misbehave.

Fan's game

| Tough | Tolerant | |
|---|---|---|
| (1,5) | (3,6) | Peaceful |
| (2,3) | **(6,4)** | Fake Hooligan |
| **(4,2)** | (5,1) | Real Hooligan |

Police's game

| | Intervene | Stay out |
|---|---|---|
| Acceptable | (1,3) | **(3,4)** |
| Aggressive | **(2,2)** | (4,1) |

# Hypergame Analysis: Mapping Radical Differences



Fans will feel provoked to resort to real hooliganism.

Fan's game

| Tough | Tolerant | |
|---|---|---|
| (1,5) | (3,6) | Peaceful |
| (2,3) | (6,4) | Fake Hooligan |
| **(4,2)** | (5,1) | Real Hooligan |

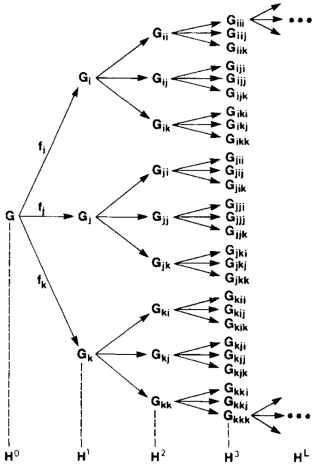
Police's game

| | Intervene | Stay out |
|---|---|---|
| Acceptable | (1,3) | (3,4) |
| Aggressive | **(2,2)** | (4,1) |

# Hierarchical Hypergames[22]



- $G_i = (N_i, P_i, A_i, O_i, V_i)$
  - $A_i \in P_i$
  - $V_i : O_i \to \mathbb{R}$, or $V_i \in \mathcal{R}(O_i)$, with $\mathcal{R}(O_i) = (\succeq_i, \succ_i, \sim_i)$
- $H^0 = G$: no perceptual differences
- $H^1 = \{H_i^0\}_{i \in N}$: misperceptions, players unaware
- $H^2 = \{H_i^1\}_{i \in N}$: at least 1 player is aware of $H^1$
- $H^3 = \{H_i^2\}_{i \in N}$: at least 1 player is aware of $H^2$



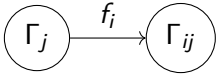

[22] M. Wang, K. W. Hipel, and N. M. Fraser, "Modeling misperceptions in games," *Behavioral Science*, vol. 33, no. 3, pp. 207–223, 1988.

# Recursive Reasoning via Game Hierarchies



A *perspective* is a player's holistic interpretation of the base conflict, composed of *images* of the game's components, generated by the perceptual mapping $f_i : \Gamma_j \rightarrow \Gamma_{ij}$, where:

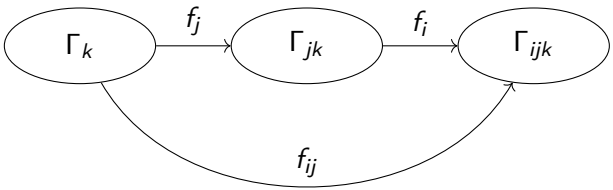
- $\Gamma_j$ represents any component from player $j$'s game
- $\Gamma_{ij}$ denotes player $i$'s interpretation of player $j$'s component;
- For any element $\gamma \in \Gamma_j$, the image $\varphi = f_i(\gamma)$ represents player $i$'s perception of that element.



# Recursive Reasoning via Game Hierarchies



When a player $i$ considers another $j$'s perception of $k$'s game, it is said to have a *higher-order expectation*, captured by a product mapping $f_i \circ f_j : \Gamma_k \rightarrow \Gamma_{ijk}$:




Then, the image $\varphi$ is the product of individual perceptual functions,
$\varphi = f_i(f_j(\gamma)) = f_i \circ f_j(\gamma) = f_{ij}(\gamma)$

# Hypergames for Multi-agent Systems

# Challenges



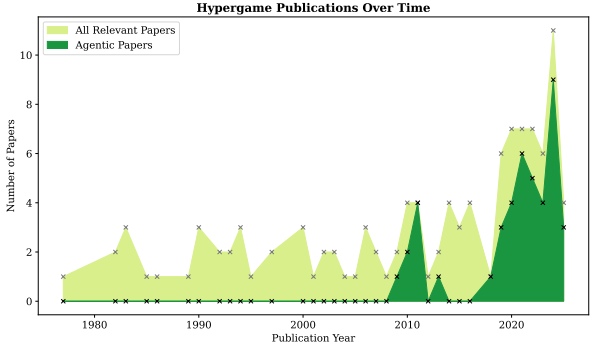
- Hypergames are designed to aid the analyst, not the decision-maker
- Defining and finding hypergame structures is non-trivial
- Hypergame Hierarchies quickly become intractable

# Hypergames in Multi-agent Systems



W.I.P survey:

- Mostly post-hoc analyses
- Lack of formal, unified language
- Limited integration
- Lacking benchmarks: no standard datasets or tasks
- Perceptual mapping remains unsolved

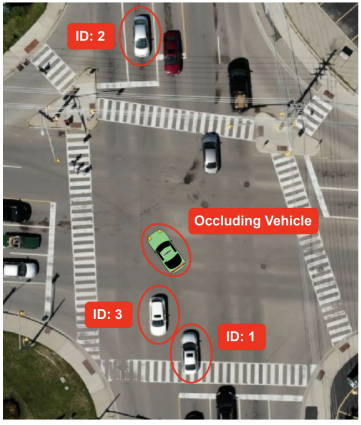


# Hypergame Traffic Management[23]



Hypergame model of occlusion:

- $H^0$: Occlusion-resolved
- $H^1$: Occlusion-naive
- $H^2$: Occlusion-aware



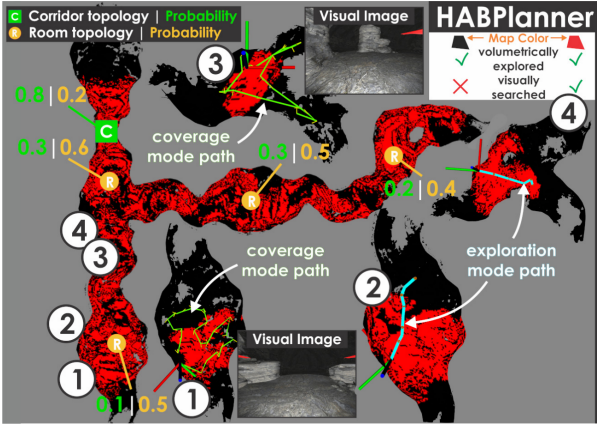

[23] M. Kahn, A. Sarkar, and K. Czarnecki, *I know you can't see me: Dynamic occlusion-aware safety validation of strategic planners for autonomous vehicles using hypergames*, 2021. arXiv: 2109.09807 [cs.RO].

# Hypergame UAV Control[24]



Hypergame vs. Nature:

- Drone builds volumetric/visual coverage
- Nature places geometric/visual challenges




[24] M. Dharmadhikari *et al.*, *Hypergame-Based Adaptive Behavior Path Planning for Combined Exploration and Visual Search*. Institute of Electrical and Electronics Engineers (IEEE), 2021, Print.

# Recursive Reasoning via Game Hierarchies

# Practical Assumptions



- Higher-order expectations $\Gamma_{ijk}$ are captured by equal-order subjective games $G_{ijk}$
- Given $\beta_i(\beta_j(\beta_k))$[25] corresponds to $i$'s beliefs about player $j$'s beliefs about player $k$'s reasoning, we assume $\beta_i(\beta_j(\beta_k)) \cong G_{ijk}$.

[25] E. Dekel and M. Siniscalchi, "Epistemic game theory," in *Handbook of game theory with economic applications*, vol. 4, Elsevier, 2015, pp. 619–702.

# Perceptual Beauty Contest Games[26]



$$G = (N, A, U, \Psi)$$

- **Players** $N = \{1, \ldots, n\}$
- **Actions** $A = A_i \times A_j, \quad A_i, A_j \subseteq \mathbb{Z}$
- **Utility Functions** $(U_i = U_j)$

$$U_i(a_i, a_{-i}) = \begin{cases} 1 & \text{if } |a_i - p\mu| < |a_{-i} - p\mu| \\ 0.5 & \text{if } |a_i - p\mu| = |a_{-i} - p\mu| \\ 0 & \text{otherwise} \end{cases}$$

- **Perspective** $\Psi = (\psi_1, \psi_2, \ldots, \psi_\kappa)$: an ordered sequence of $\kappa$ perception steps, $\psi_1$ denotes the interpreter (creator) of $\Psi$

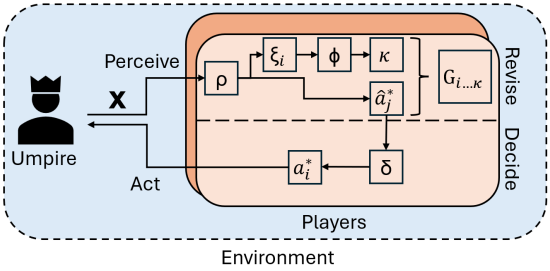
[26] V. Trencsenyi, A. Mensfelt, and K. Stathis, "Approximating human strategic reasoning with llm-enhanced recursive reasoners leveraging multi-agent hypergames," *arXiv preprint arXiv:2502.07443*, 2025.

# Multi-agent Centralized Hypergames



The umpire $\upsilon$ is a pseudo-player:

- Defines the **true game** $G^* = (N, A, U, \Psi = (\upsilon))$
- Facilitates games
- Validates player actions
- Translates between subjective games



# Multi-agent Centralized Hypergames

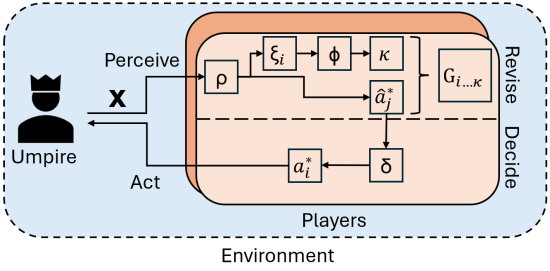


**Revision:**

- **Reason** $\rho : X \to \Xi, \mathbb{Z}$
- **Evaluate** $\phi : \Xi \to \mathbb{Z}^+$

**Decision:**

- $\delta : \mathbb{Z} \to A_i$




$G_{i...\kappa} = (\{i, j\}, A, U, \Psi = (\psi_i, \ldots, \psi_\kappa))$ captures the player's beliefs used for reasoning.

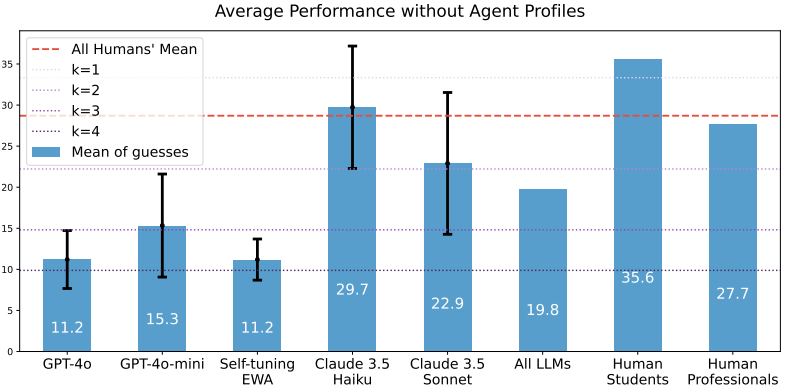
# k-levels and $\kappa$-levels



- k-levels are solely derived from players' numerical choices $a_k = p^k \cdot a_0$
  - Hard to distinguish noise from sophistication
- $\kappa$ provides a more informed, semantically-backed measure to k-levels
  - It is still subject to semantic errors $\rightarrow$ semantic validation !

# Matching Overall Human Reasoning Level

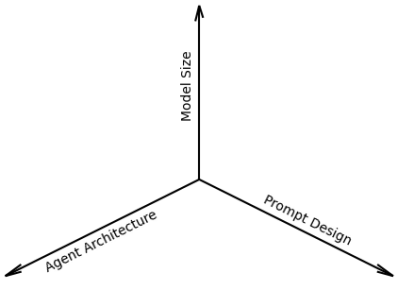




# The Influence of Human-inspired Architectural Complexities[27]



- **Model size:** Haiku, Sonnet
- **Architecture:** One-shot prompt → Decoupled reasoning
- **Prompt Model:** ∅ →Context → MoA




Agentic sophistication

[27] V. Trencsenyi, A. Mensfelt, and K. Stathis, *The influence of human-inspired agentic sophistication in llm-driven strategic reasoners*, 2025. arXiv: 2505.09396 [cs.AI].

# Agent Profiles



**Simple Profile**

You are an economics professor, with expert level domain knowledge in game theory.

**Bio**

Dr. VolumetricBrin is a distinguished economics professor ... at the Wharton School of the University of Pennsylvania. Born in Boston, Massachusetts in 1975, Brin demonstrated an early aptitude for mathematical reasoning and strategic thinking. ... Dr. Brin is known for his interdisciplinary approach, often bridging computational theory, behavioral economics, and strategic analysis in his research and teaching.

# A Model of Appropriateness[28]



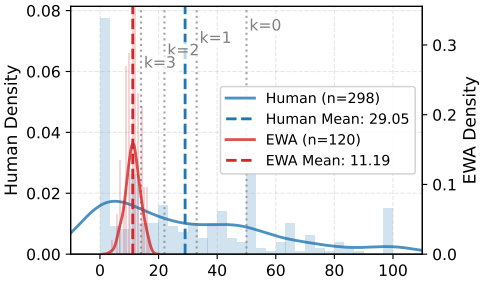
1. What kind of situation is this?
2. What kind of person am I?
3. What should a person like me do in a situation like this?

[28] A. S. Vezhnevets, J. P. Agapiou, A. Aharon, *et al.*, "Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia," *arXiv preprint arXiv:2312.03664*, 2023.

# LLM-driven Recursive Reasoners Towards Human-like Strategic Reasoning

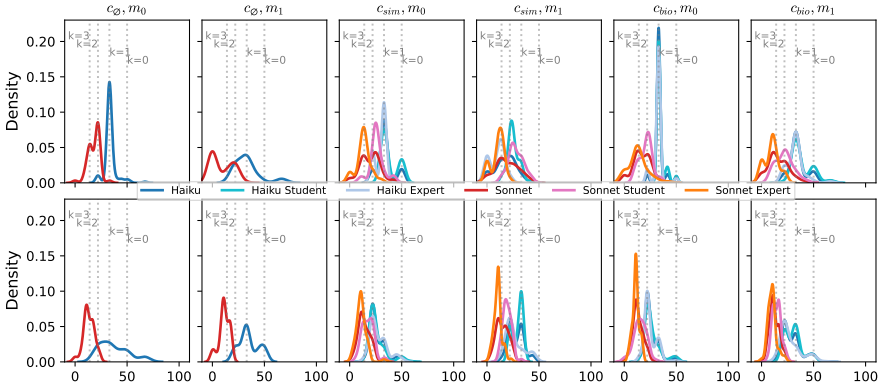


- Analysis of population-wise and role-specific performance
  - Average k-level sophistication
  - Matching human guess distribution
  - Frequency of zero guesses
- Out-of-sample validation
  - Obfuscated game description
  - Shifted guess interval



# Matching Distribution of Human Choices

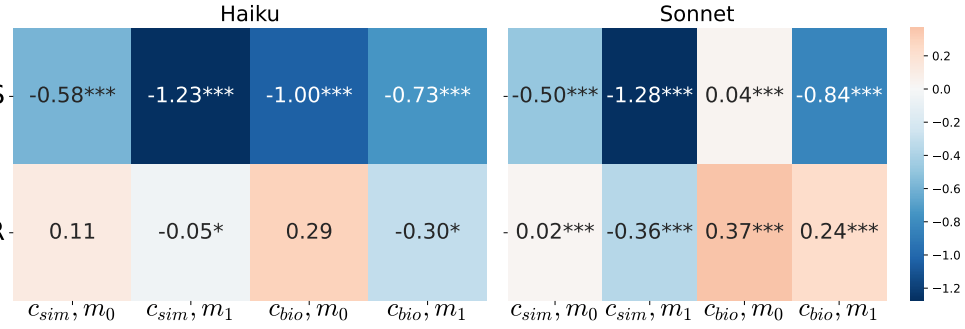





Smaller, simpler agents approximate human performance better.

# Human/Agent Cohort-wise Mean k-level Δ

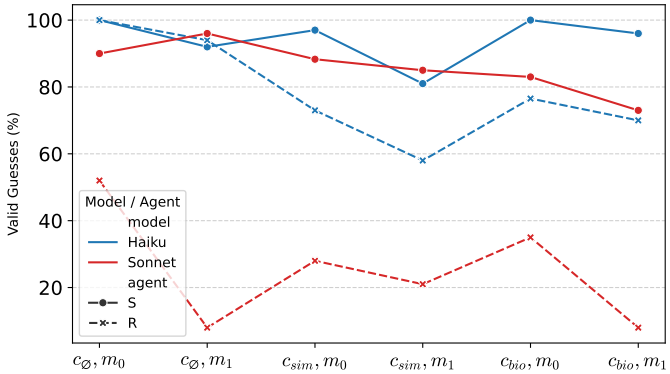





The smaller model wrapped in complex architecture, using simpler prompts replicate human bounded rationality closely.

# Out of Sample Validation






Simpler models are more robust, the inclusion of MoA hurts results in each case.

# LLM-as-agent Pointers



No One-Size-Fits-All:

- Agentic design should be guided by the task, not by convenience or theoretical purity.
- Wrapping LLMs in structured agentic workflows enables deeper behavioral analysis and better control than unstructured LLM-as-agent setups.

# Overall Conclusions

# Conclusions I

## Formalizing Strategic Interactions



- Many $n \times n$ games from game theory effectively capture key social interactions.
- These can be formalized in a **GDL-like language** for interpretable player reasoning.
- This reasoning enables agents to analyze strategies and tactics in social and strategic settings.
- However, manual formalization can be time-consuming:
  - We propose **autoformalization**, where agents receive interactions in natural language and use LLMs with in-context learning to formalise games.
  - This works well for a class of $n \times n$ games.

# Conclusions II

## Autoformalization in Strategic AI



- A lightweight GDL-based formalism based on logic-programming solver supports:
  - Serving as the target language for autoformalized games.
  - In-context specification for LLM-based formalization.
  - Agents reasoning and experimenting with formalized games.
- The framework enables experimentation and evaluation of autoformalization workflows.
- For more complex games and interactions, extensions are needed - e.g., **temporal reasoning over events** (as in the Event Calculus).

# Conclusions III

## Towards Human-Like Strategic Reasoning



- Standard game-theoretic assumptions (e.g., perfect rationality, complete information) often fail in realistic social scenarios.
- Multi-agent systems display strategic complexity beyond classical models:
  - Diverse, sometimes conflicting goals.
  - Partial observability and hidden intentions.
  - Dynamic and evolving interactions.
- Game representations should therefore be:
  - **Formal** — interpretable using logical descriptions.
  - **Expressive** — capturing nuanced social and strategic behaviours.
  - **Computationally usable** — enabling automated agents and experiments.
- Our framework paves the way for bridging game theory and multi-agent systems via autoformalization.