21. Recognition of Non-Cooperative Behavior

University of Groningen, contact: j.d.top@rug.nl

People

J.D. Top
PhD Candidate

prof. dr. Verbrugge
First supervisor

prof. dr. Jonker
Second supervisor

dr. De Weerd
Daily supervisor

Outline

- Investigate the logical and computational foundations of deception and deception detection in hybrid groups.
- Lay the theoretical groundwork for modelling and analyzing non-cooperative behavior in several communicative contexts such as negotiation games and coalition formation games.
- Develop principled methods for the design of software agents that can detect when other group members are engaged in non-cooperative behavior such as lying.
- Build agent-based models and/or computational cognitive models of deception and deception detection.
- Use simulation experiments in order to predict the outcomes of lab experiments to be performed.

First article (overview)

'Predictive Theory of Mind Models Based on Public Announcement Logic'

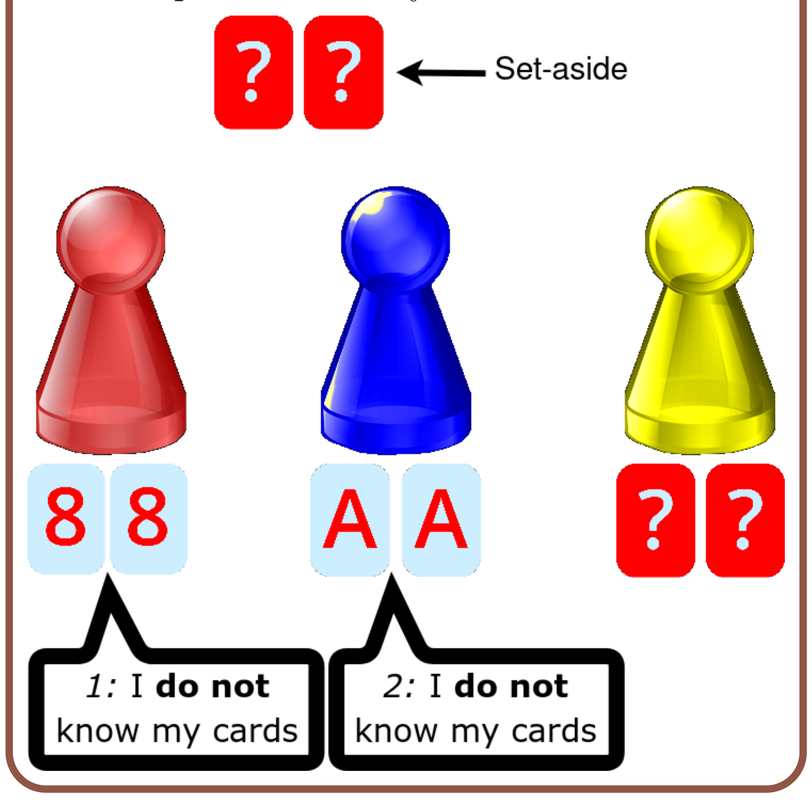
- Presents a logic for Theory of Mind (ToM)
- Creates computational models of agents using this logic
- Performs model fitting of these models on the data of Cedegao et al. (2021), where participants play **Aces and Eights**

The ability to attribute mental states to others, such as beliefs, desires, intentions... Can be used recursively: '*you know that I know...*'

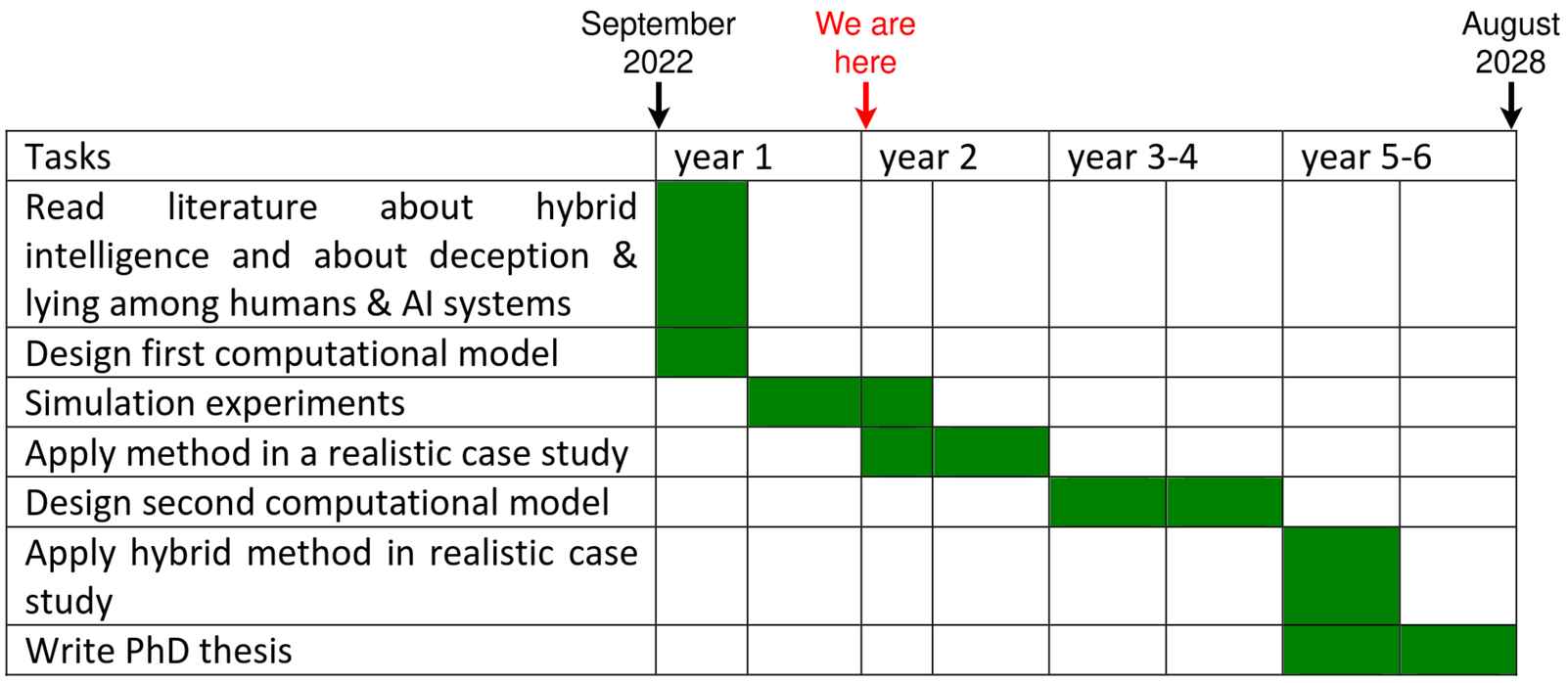
First article (setting)

Aces and Eights:

- Played with three players
- Uses a deck of four Aces and four Eights
- Each player gets two cards
- You can only see other players' cards
- You have to announce *whether* you know what your cards are
- Example: what will yellow answer?



Next

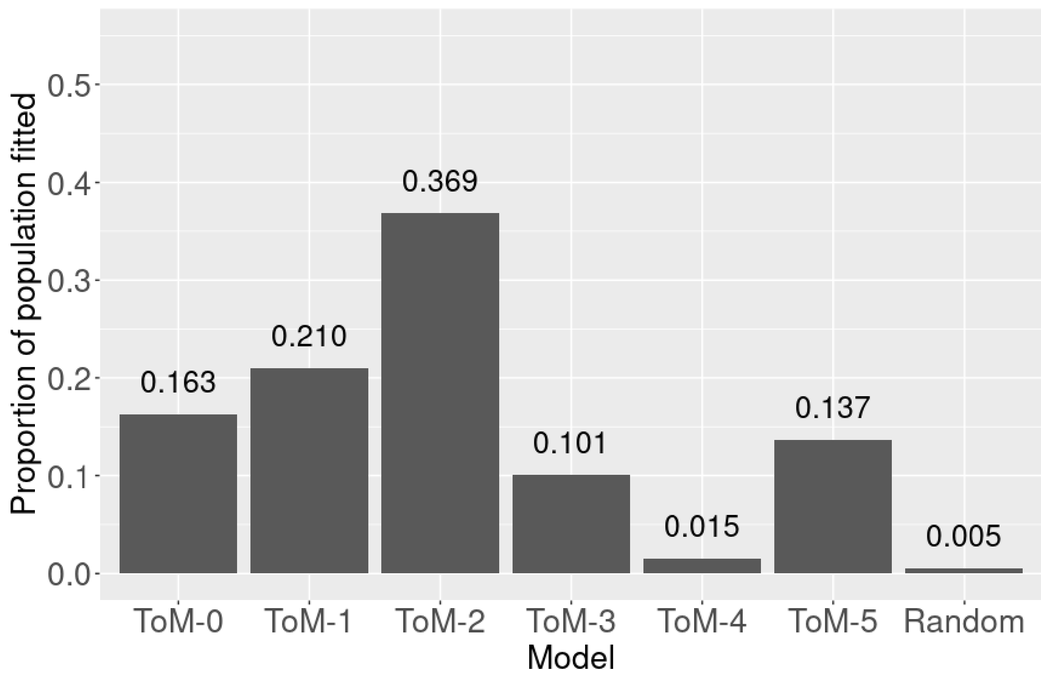


Tasks	year 1		year 2		year 3-4		year 5-6	
Read literature about hybrid intelligence and about deception & lying among humans & AI systems	■							
Design first computational model	■							
Simulation experiments		■	■					
Apply method in a realistic case study			■	■				
Design second computational model					■	■		
Apply hybrid method in realistic case study							■	
Write PhD thesis							■	■

- Special six-year PhD position with 60% research and 40% teaching
- Next, create a formal logic that can combine Theory of Mind and lying (in progress)
- Then, apply this logic in a realistic case study

First article (results)

- The plot shows our estimated frequencies of ToM levels in Cedegao's data
- ToM-n means you can switch perspectives no more than n times, e.g. '*I know that you know that I know my cards*' is two switches
- Crucially, we assume there is *no limit* on reasoning about your *own* knowledge
- The peak at ToM-2 is comparable to previous results and validates our logic.
- Despite being deterministic, our models predict ≈75% of participant answers



References

Cedegao, Z., Ham, H., Holliday, W.H.: Does Amy know Ben knows you know your cards? A computational model of higher-order epistemic reasoning. In: *Proceedings of the 43th Annual Meeting of the Cognitive Science Society*. pp. 2588–2594 (2021)

Top, J.D., Jonker, C.M., Verbrugge, L.C., De Weerd, H.A.: Predictive Theory of Mind Models Based on Public Announcement Logic. In: *Preproceedings of the DaLí 2023 International Workshop*. pp. 40-57 (2023)