Cooperation, Deception, and Theory of Mind in a Cyclic Game with Inter-Player Signalling

Jakob Dirk Top[1]✉, Harmen de Weerd[1], Abisharan Raveenthiran[1], Catholijn Jonker[23], Rineke Verbrugge[1]

[1]Bernoulli Institute, University of Groningen, Netherlands
[2]Interactive Intelligence, Delft University of Technology, Netherlands
[3]LIACS, Leiden University, Netherlands

✉{j.d.top@rug.nl}

university of groningen

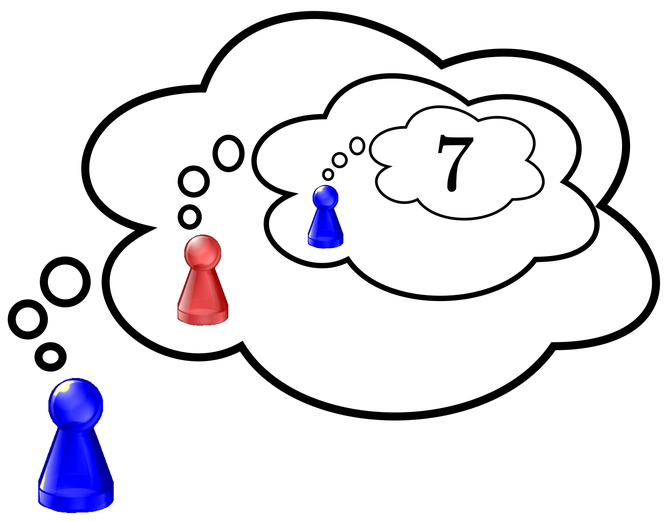
Hybrid Intelligence

Main Goal

- To study **cooperation**, **deception**, and **theory of mind** using our novel **Mod-Signal game**.

Theory of Mind

- The ability to **reason about mental states of others**.
- Your theory of mind **order** is the **number** of perspective switches you use.
- For example: In "*Blue* ***thinks*** *that Red* ***thinks*** *that Blue thinks about playing 7*", Blue uses **two** perspective switches, so he uses **second**-order theory of mind.
- Humans usually use **no more than second-order** theory of mind.



Experiment

- **21 participants** played the Mod-Signal game with **one another**, and with a **computer agent**.
- The **computer agent** would **copy** the participant's behaviour: It cooperates when the participant cooperates, and competes when the participant competes.

Results (cooperation)

- Surprisingly, participants mainly played **non-cooperatively**, despite cooperation being **faster** and yielding **more points**.
- Perhaps this was caused by participants receiving **fixed payment amounts**.
  - In future work, participants should be paid based on the points they gain.

References

[1] de Weerd, H., Verbrugge, R. and Verheij, B. (2013), How much does it help to know what she knows you know? An agent-based simulation study. *Artificial Intelligence*, 199-200:67-92. doi:10.1016/j.artint.2013.05.004.

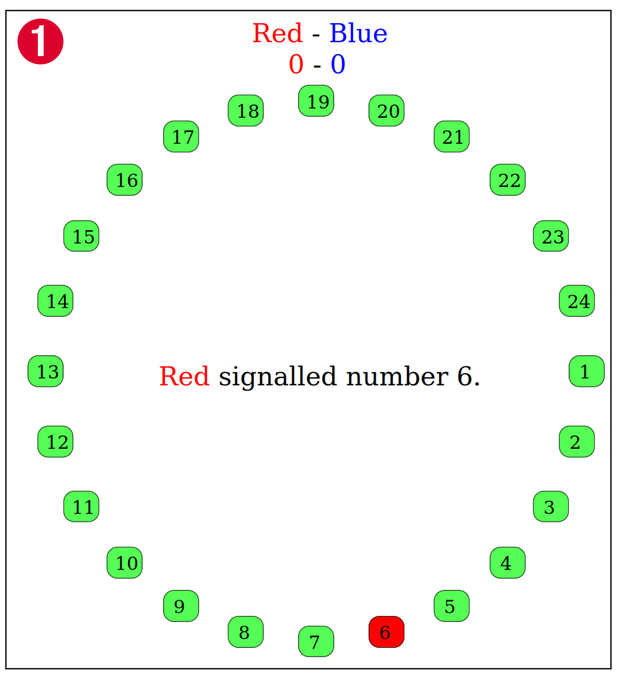
[2] van Ditmarsch, H., Hendriks, P. and Verbrugge, R. (2020), Editors' review and introduction: Lying in logic, language, and cognition. *Topics in Cognitive Science*, 12:466-484. doi:10.1111/tops.12492

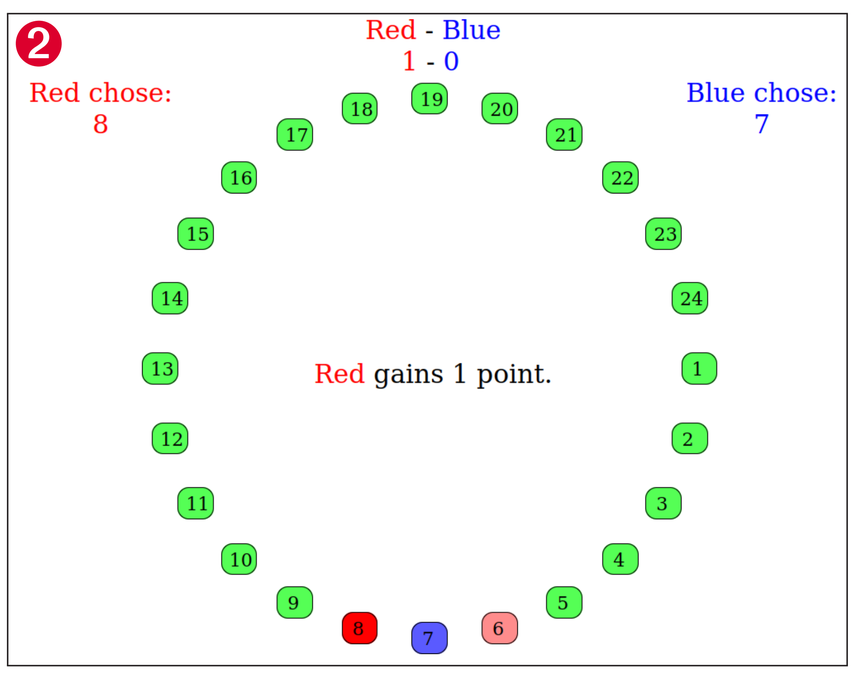
[3] Frey S. and Goldstone R.L. (2013), Cyclic game dynamics driven by iterated reasoning. *PLOS One*, 8(2):e56416. doi:10.1371/journal.pone.0056416

The Mod-Signal Game

- Each round, one player is the **signaller**, and the other player is the **receiver**.
- The **signaller** first signals one of 24 actions placed on a circle.
- Then, **both players** simultaneously select an action.
- If your action is exactly one space ahead of the other player's action, you gain a point.

Example





❶ First, Red signals that she will play 6.

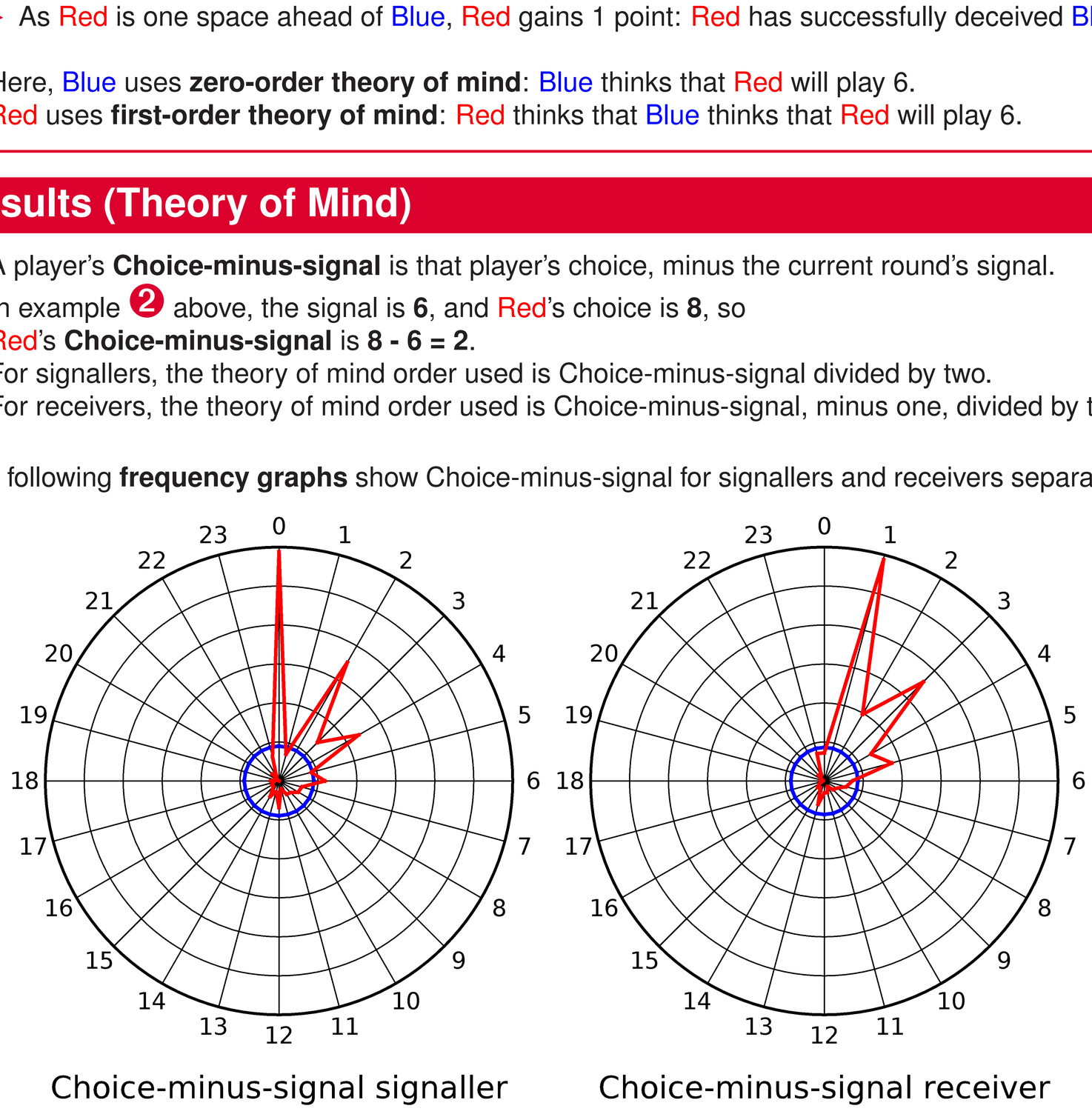
❷ Then, Red and Blue **simultaneously** choose an action.
- Blue plays 7, hoping that Red signalled truthfully, such that he can gain a point.
- Red plays 8, hoping that Blue believes her signal and plays 7.
- As Red is one space ahead of Blue, Red gains 1 point: Red has successfully deceived Blue.

- Here, Blue uses **zero-order theory of mind**: Blue thinks that Red will play 6.
- Red uses **first-order theory of mind**: Red thinks that Blue thinks that Red will play 6.

Results (Theory of Mind)

- A player's **Choice-minus-signal** is that player's choice, minus the current round's signal.
- In example ❷ above, the signal is **6**, and Red's choice is **8**, so Red's **Choice-minus-signal** is **8 - 6 = 2**.
- For signallers, the theory of mind order used is Choice-minus-signal divided by two.
- For receivers, the theory of mind order used is Choice-minus-signal, minus one, divided by two.

The following **frequency graphs** show Choice-minus-signal for signallers and receivers separately:



Choice-minus-signal signaller

Choice-minus-signal receiver

- The **peaks** correspond to different theory of mind orders.
- It can be seen that participants used **no more than second-order theory of mind**.
- Furthermore, **lower theory of mind orders are used more frequently**.