Partner Modelling Emerges in Recurrent Agents (But Only When it Matters)

Ruaridh Mon-Williams, Max Taylor-Davies, Elizabeth Mieczkowski, Natalia Vélez,
Neil R. Bramley, Yanwei Wang, Thomas L. Griffiths, Christopher G. Lucas

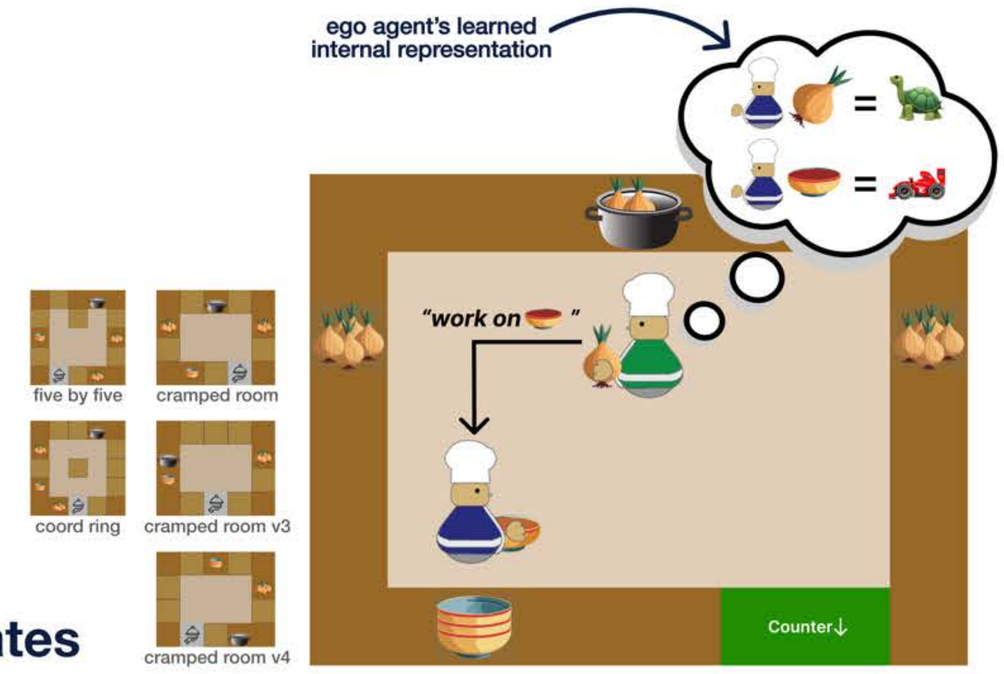
Can parter modelling emerge spontaneously from the pressure of flexible collaboration, without dedicated architectures or auxiliary objectives?

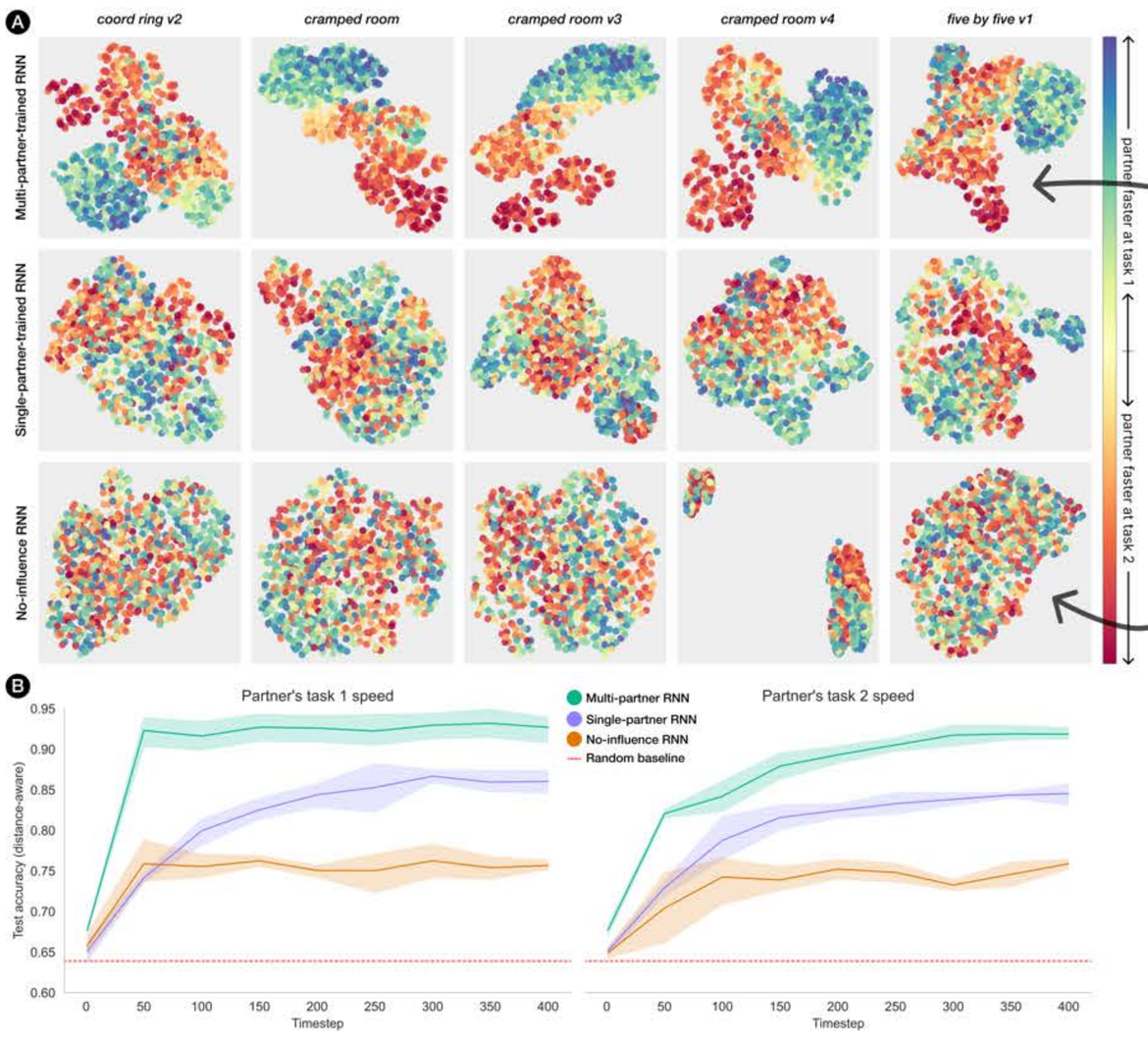
Experimental setup

We use **Overcooked** to study flexible cooperation, where a pair of agents must produce soup by coordinating **two subtasks** (cooking ingredients and serving).

An RNN ego agent is trained alongside a population of partners with **different subtask speeds**, and has the ability to **control which subtask** their partner works on.

After training, we pair the ego agent with partners from an **unseen test distribution** + analyse its **internal hidden states**

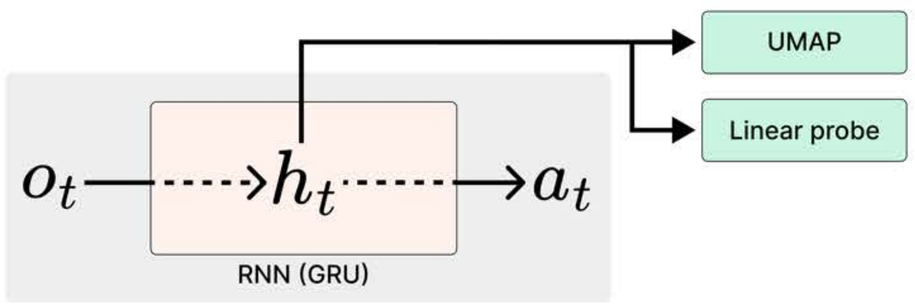




Ability to influence partner task drives representation

The RNN ego agent develops **structured representations** that **encode partners' abilities** across the two subtasks

If we train the ego agent **without** the ability to control subtask allocation, the emergence of these representations is **strongly impaired**

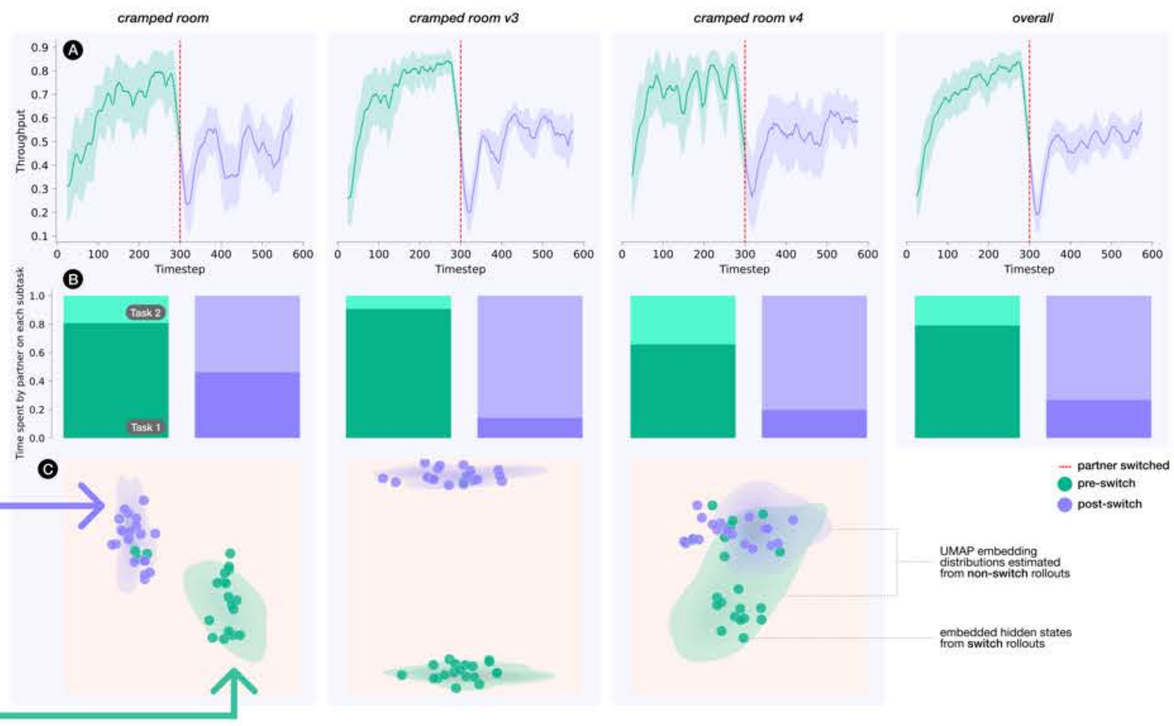


Representations adapt online to new partners

In a second experiment, we **switch** the ego agent's partner halfway through the episode

The **first partner** is fast at task 1 and slow task 2; the **second** is slow at task 1 and fast at task 2

We find that the ego agent's hidden states recorded **pre-** and **post-**switch **reflect the change in their partner's properties**



Check out the paper for:

- ★ Confirmatory results in a second task environment (CoinGame)
- ★ Additional analysis of RNN representations
- ★ More pretty figures!

ARXIV