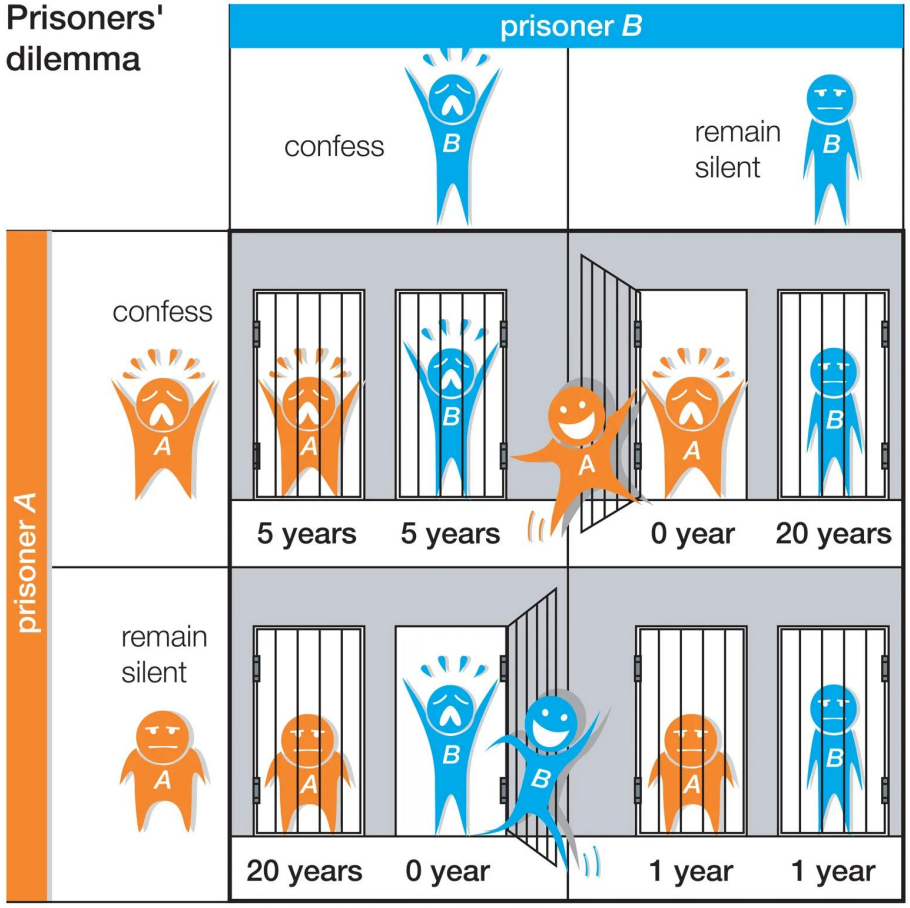
Empathy paper knowledge transfer

Sanjeev Namjoshi
12/16/25

Aims for active inference model

- Cooperation as emergent behavior through empathy
- Infer beliefs of other agent in response to your actions (theory of mind)
- Factor in emotional state of other agent (via valence/arousal calculated from VFE/EFE)
- Empathy parameter controlling degree to which other agent's beliefs are used for decision-making



Source: https://www.britannica.com/science/game-theory/The-prisoners-dilemma

An analytical model of active inference in the Iterated Prisoner's Dilemma

Daphne Demekas[*1,2], Conor Heins[†1,3,4,5], and Brennan Klein[‡1,5,6]

Demakes et al. 2023

Generative model

Variable Name	**Notation**
Hidden States	$\mathbf{s} \in \{\text{CC}, \text{CD}, \text{DC}, \text{DD}\}$
Observations	$\mathbf{o} \in \{\text{CC}, \text{CD}, \text{DC}, \text{DD}\}$
Actions	$\mathbf{u} \in \{u^C, u^D\}$
Observation Model	$P(\mathbf{o}_t \vert \mathbf{s}_t; A) = Cat(\mathbf{A})$
Transition Model	$P(\mathbf{s}_{t+1} \vert \mathbf{s}_{t-1}, \mathbf{u}_{t-1}; B) = Cat(\mathbf{B})$
Transition Model Parameter	$P(B) = \prod_{ju} P(B_{\bullet ju}), \quad P(B_{\bullet ju}) = Dir(\mathbf{b}_{\bullet ju})$
Initial State Prior	$P(\mathbf{s}_1; D) = Cat(\mathbf{D})$
'Biased' State Prior (Reward)	$\tilde{P}(\mathbf{s}; C) = Cat(\mathbf{C}), \text{ s.t. } \ln \mathbf{C} = [3, 1, 4, 2]$

Table 2: Generative model variables and notation.

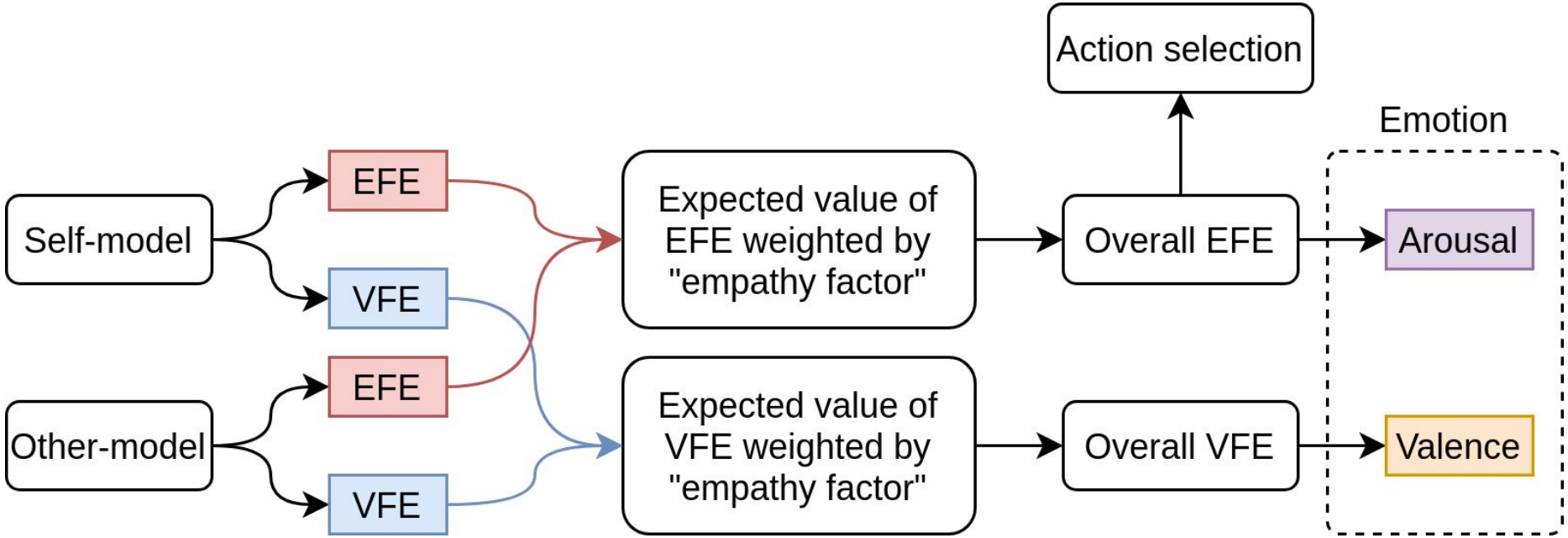
Demakes et al. 2023

Payoff matrix

		Player 2	
		Cooperate (**C**)	Defect (**D**)
Player 1	**C**	$(3, 3)$	$(1, 4)$
	D	$(4, 1)$	$(2, 2)$

Demakes et al. 2023

Active inference model (current form)



Pattisapu et al. 2024

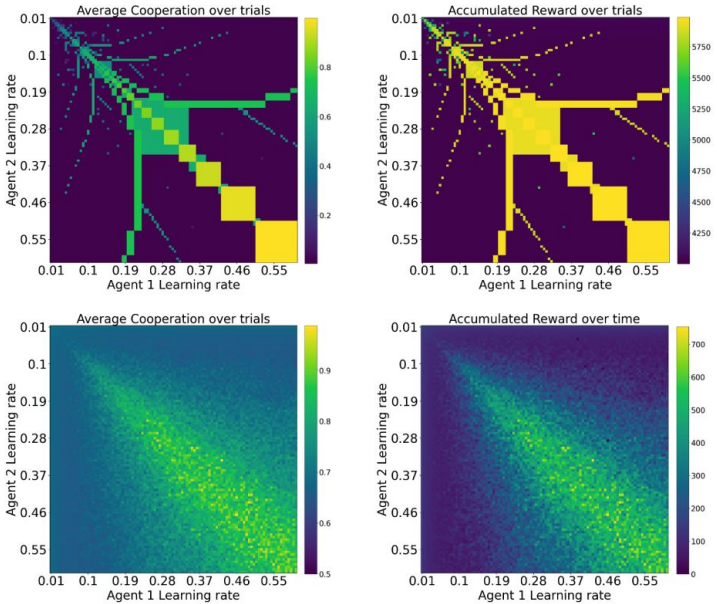
- Stochastic A matrix
- Learning for B matrix enabled
- Learning for C matrix (maybe)

Active inference model - major goals

- Grid search over empathy parameter setting
- Find the conditions under which cooperation emerges

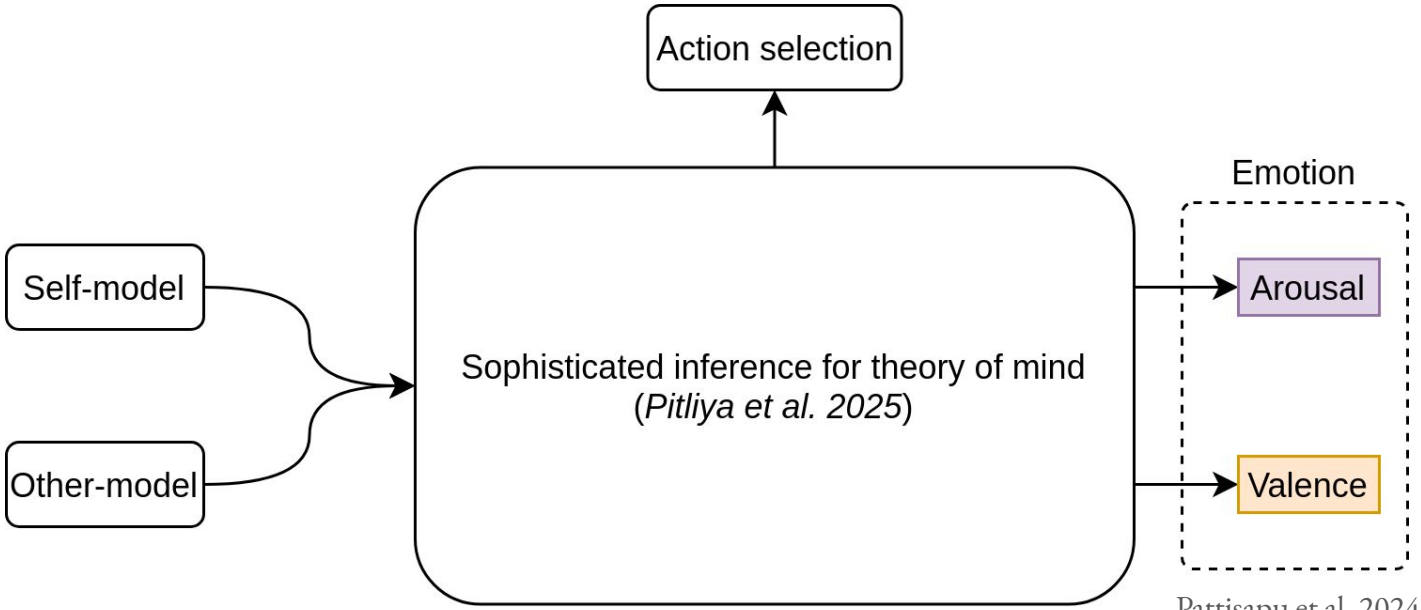
Future directions:

- Examine emotional state over time
- Compare results with and without B learning
- Examine how preference learning is affected by the empathy parameter
- Non-identity A matrix
- Expansion to n-player game



Demakes et al. 2023

Active inference model (revised)



Pattisapu et al. 2024

Active inference model (revised)

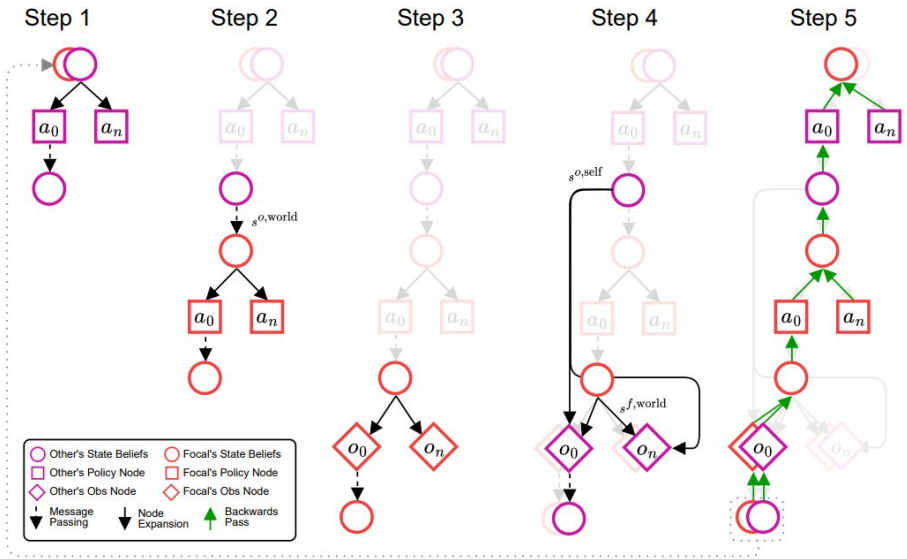


Fig. 1: **Recursive Planning Tree for Theory of Mind.** Red and purple represent the focal and other agent's nodes respectively. Circles indicate the agent's beliefs, squares indicate evaluated actions and diamonds indicated expected observations. For a detailed description of each step, see Section 2.3.

Pitliya et al. 2025

Active inference model (revised empathy factor)

$$G(o^f_\tau, o^o_\tau, a^f_\tau, a^o_\tau) = \mathbb{E}_{Q(o^f_{\tau+1}, o^o_{\tau+1} | a^f_{\leq\tau}, a^o_{\leq\tau})} \left[-\ln P(o^f_{\tau+1}|C^f) - \mathbb{D}_{\mathrm{KL}}[Q(s^f_{\tau+1}|o^f_{\tau+1}) || Q(s^f_{\tau+1})] \right] + \mathbb{E}_{Q(a^f_{\tau+1}|o^f_{\tau+1}) Q(a^o_{\tau+1}|o^o_{\tau+1}) Q(o^f_{\tau+1}, o^o_{\tau+1}|a^f_{\leq\tau}, a^o_{\leq\tau})} \left[G(o^f_{\tau+1}, o^o_{\tau+1}, a^f_{\tau+1}, a^o_{\tau+1}) \right] \quad (2)$$

- *Option 1*: Encode empathy factor by weighting EFE in the recursive expansion based on contribution from focal or other agent
- *Option 2*: Encode empathy factor through precision on observation/action expansion

Pitliya et al. 2025

Scoping options/difficulties

- Access to theory of mind code from Pitliya et al. 2025?
- Stop at emergence of cooperation under simplest scenario?
  - 2 player game, iterated
  - Empathy parameter
  - Identity A matrix (KL-control)
  - B matrix learning
  - No emotion component analysis
- Add any of the following?
  - Examine emotional state over time
  - Compare results with and without B learning
  - Examine how preference learning is affected by the empathy parameter
  - Non-identity A matrix
  - Expansion to n-player game

References

- **Demakes D, Heins C, and Klein B**. 2023. An analytical model of active inference in the Iterated Prisoner's Dilemma. arXiv:2306.15494v2
- **Pattisapu C, Verbelen T, Pitliya RJ, Kiefer AB, Albarracin M**. 2024. Free Energy in a Circumplex Model of Emotion. arXiv:2407.02474v1
- **Pitliya RJ, Catal O, Van de Maele T, Pezzato C, Verbelen T. 2025**. Theory of Mind Using Active Inference: A Framework for Multi-Agent Cooperation. arXiv:2508.00401v2