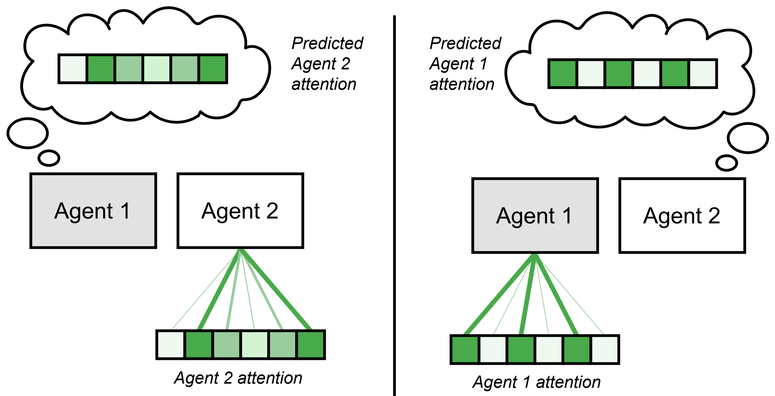
Can we build more cooperative deep learning models from theories of human cognition?

PRINCETON Neuroscience INSTITUTE

Kathryn Farrell[1], Kirsten Ziman[2], Michael Graziano[1,2]

[1]Princeton Neuroscience Institute
[2]Department of Psychology
Princeton University, Princeton NJ, 08544

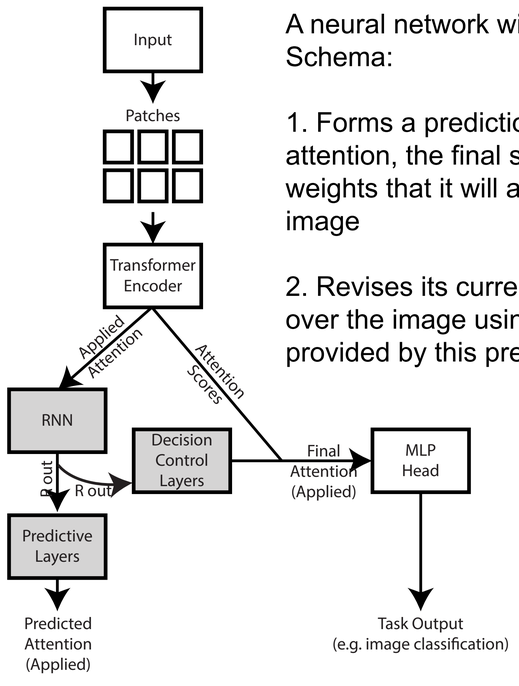
I. What allows an agent to cooperate with others?



Predictive. A socially cooperative agent should predict the attentional states (and thus, strategies) of others.

Predictable. A socially cooperative agent should have a regular attentional state that is predictable to others.

II. The "Attention Schema Theory" proposes attentional self-modeling.



A neural network with an Attention Schema:

1. Forms a prediction of its own attention, the final significance weights that it will apply over an image

2. Revises its current attention over the image using the context provided by this prediction.

III. Results: Classifying attention tensors

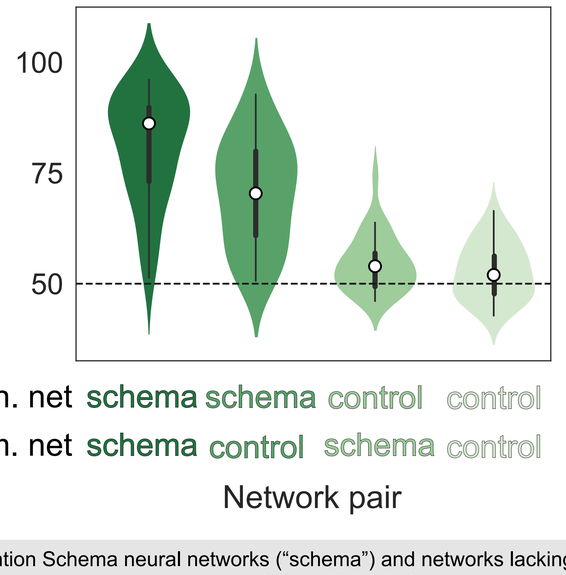


Figure 1. Attention Schema neural networks ("schema") and networks lacking the additional prediction and revision modules ("control") classified images which contained visualizations of the attention generated by other neural networks of both architectures. Schema nets significantly outperform control networks in the classification of attention visualizations from other neural networks. Both schema and control networks perform better when classifying visualizations of the attention generated by schema nets than when classifying the attention generated by control nets.

IV. Results: Nonspecific image tasks

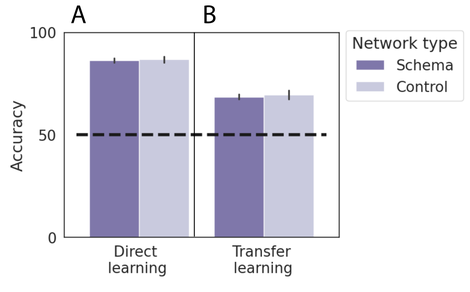
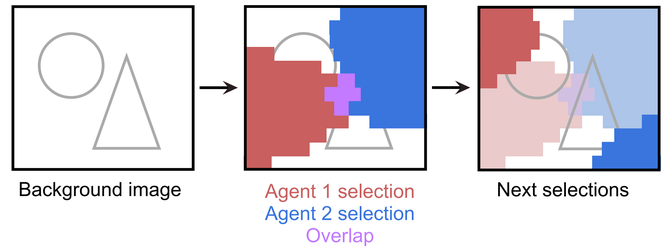


Figure 2. No advantage is observed across the schema and control architectures when they learn other binary image classifications (e.g., golf ball versus garbage truck).

V. Results: "Painting" over an image together



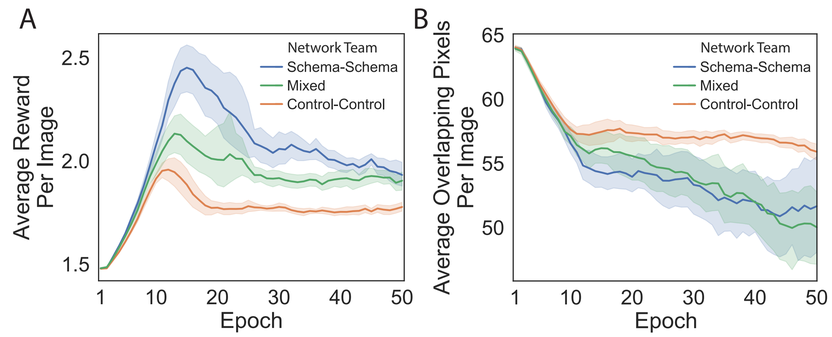


Figure 4. Pairs of schema networks, pairs of control networks, and schema-control pairs ("Mixed") collaborating on a visual task: jointly selecting all pixels in an image while minimizing overlapping pixels between their selections. A: Schema-schema teams achieved the highest episode mean rewards. B: Teams with at least one schema net outperformed control-control teams in minimizing the overlap in their selections.

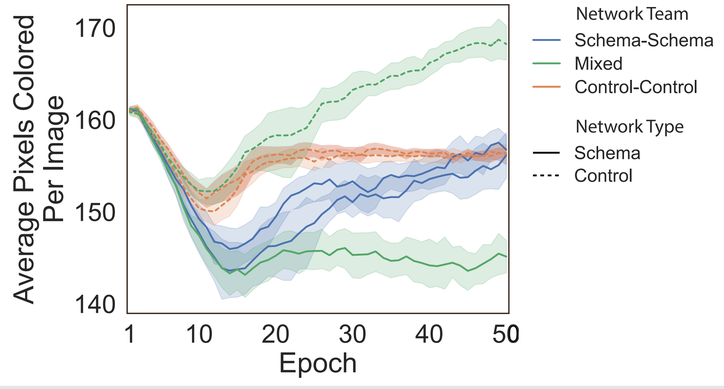


Figure 5. Schema nets (solid lines) adopt more conservative pixel selection strategies than control nets (dotted lines). The difference is especially stark within Mixed teams, when the schema network may be learning to avoid a control network with poor collaborative abilities.