Capturing Intelligence in Brains and Machines: A Long-Term Project

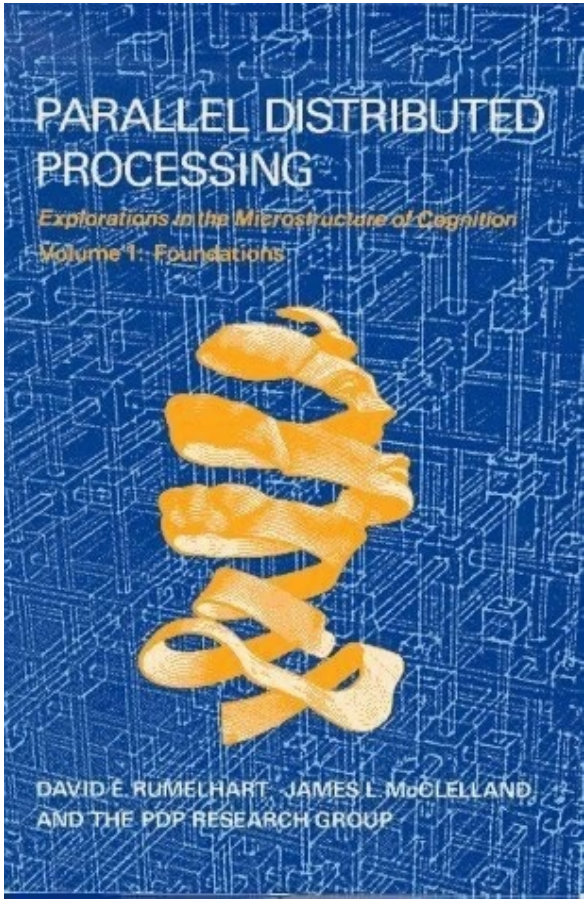
Jay McClelland
Stanford University and Google DeepMind

My points today

- For scientists, models are not the final word, or instantiations of belief.
- They are *tools for exploring the implications of ideas.*
  - When a model goes beyond what has been done before, is sharpens our observations, allowing us to see where our ideas succeed and where they fail.
- Like in Physics and Biology, our models in Cognitive Neuroscience today can seem powerful, but they remain incomplete.
- I invite the thought that this state of affairs may be ongoing.
- Investment in research and development will continue to pay off over the very long term.

What sort of a system do we need to model intelligence?

- In the 1980's we argued that the AI systems of the day would never work.
- We proposed a set of ideas called *Parallel Distributed Processing* as an alternative.
- At the time, they generated a lot of interest, but they had limitations, leading them to be abandoned around the turn of the century.
- Now, AI models are driven by the principles outlined in *PDP*



Principles of PDP used in AI systems (AKA Deep Neural Networks, LLM's, …)

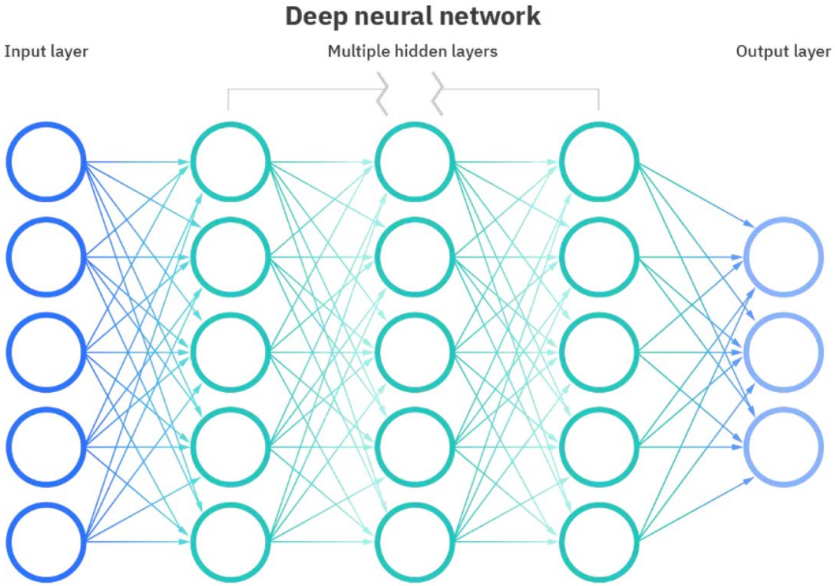
Processing occurs via propagation of graded activation signals among simple processing units

Representations are distributed patterns of activation across these units

The knowledge in the system is in the connections

Learning occurs through adjusting connections to:

- Match a target output
  - Image label, next word in a sequence
- Or maximize reward



In some ways, today's AI systems based on PDP ideas are pretty smart!

- They function well enough to be in use everyday for speech recognition, language translation, and many other things
- They exceed human performance in games like Go and Chess, and they are getting more and more powerful everyday.
- In some ways, they have begun to exceed our capabilities!

Large language models: are they on the verge of capturing intelligence?

Sam Altman, CEO, OpenAI

Sam Altman @sama · Nov 30

Replying to @sama

language interfaces are going to be a big deal, i think. talk to the computer (voice or text) and get what you want, for increasingly complex definitions of "want"!

this is an early demo of what's possible (still a lot of limitations--it's very much a research release).

18 60 884

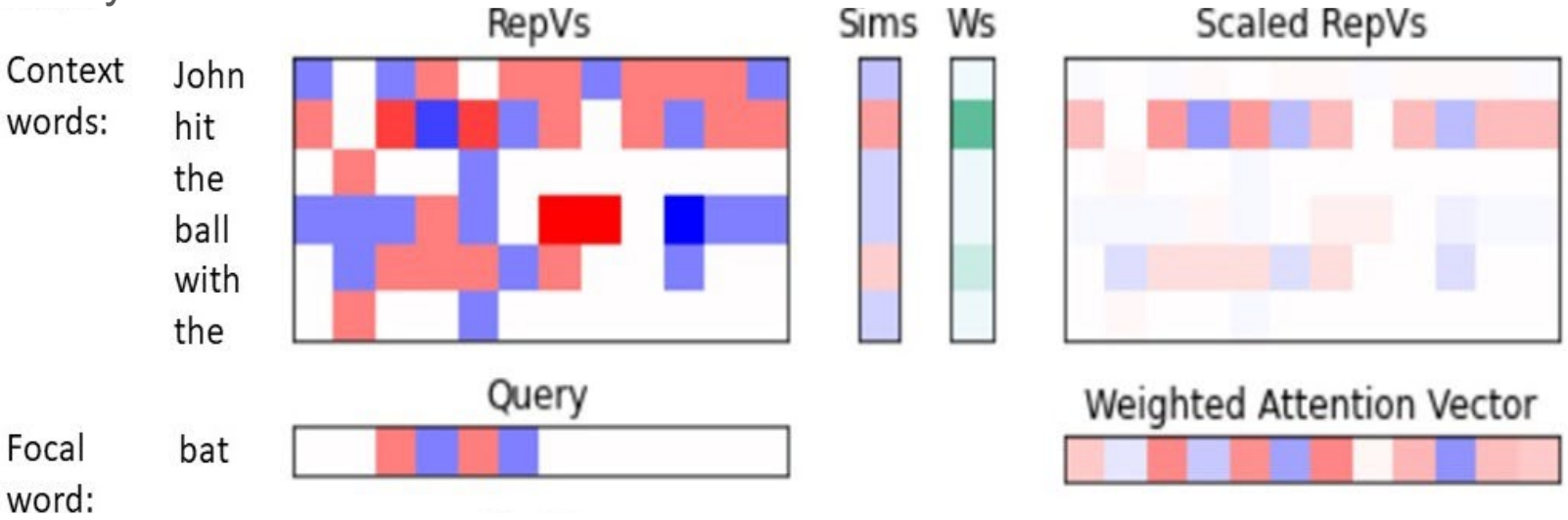
Sam Altman @sama · Nov 30

soon you will be able to have helpful assistants that talk to you, answer questions, and give advice. later you can have something that goes off and does tasks for you. eventually you can have something that goes off and discovers new knowledge for you.

28 123 1,091

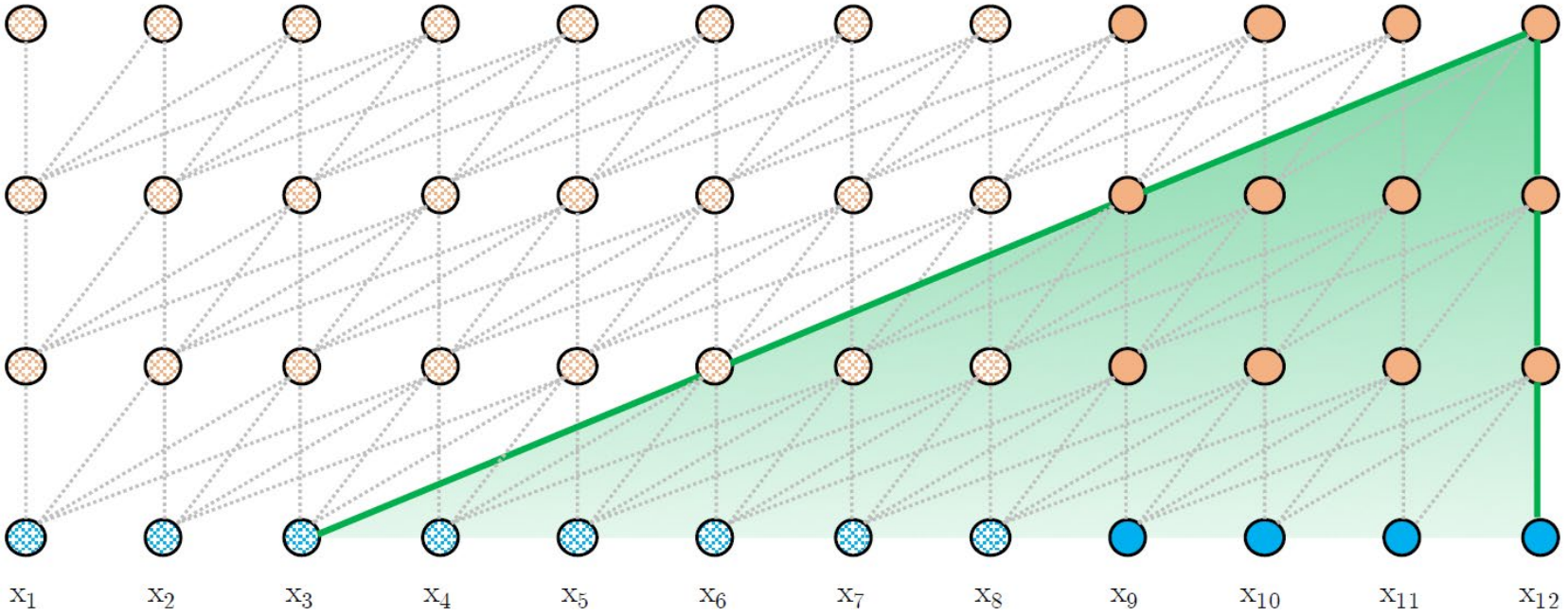
A Major Innovation in Neural Networks

Query Based Attention



$$\mathbf{a} = \sum_c w_c (\mathbf{r}_c \cdot \mathbf{q})\, \mathbf{r}_c$$

Model width and depth: GPT-3 series up to 4k x 96
GPT-4 up to 32k x ???



My points today

- For scientists, models are not the final word, or instantiations of belief.
- They are *tools for exploring the implications of ideas.*
  - When a model goes beyond what has been done before, is sharpens our observations, allowing us to see where our ideas succeed and where they fail.
- Like in Physics and Biology, our models in Cognitive Neuroscience today can seem powerful, but they remain incomplete.
- I invite the thought that this state of affairs may be ongoing.
- Investment in research and development will continue to pay off over the very long term.

Case Study #1: Bonnen, Yamins & Wagner, *Neuron,* 2021

- CNN models of the ventral visual stream account well for neural data but not for all aspects of human perceptual abilities
  - Networks exhibit a ‘Texture Bias’ while humans see shape

(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

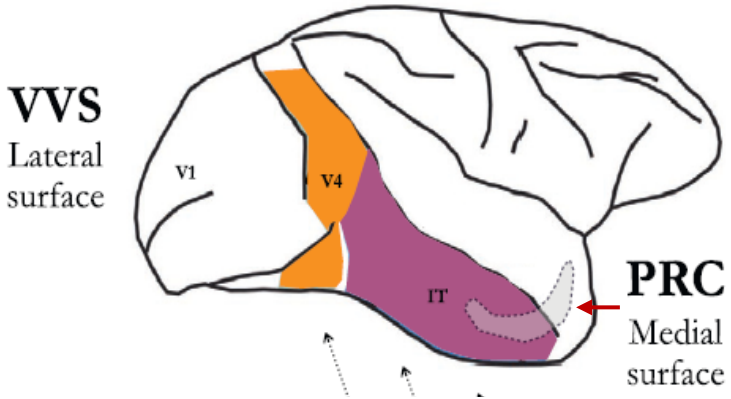
71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

(c) Texture-shape cue conflict

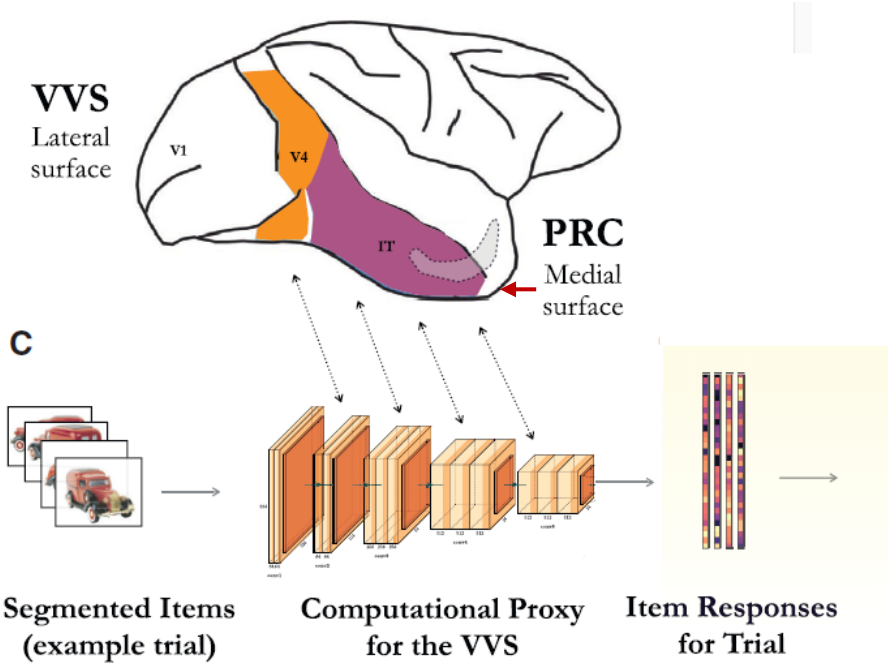
63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

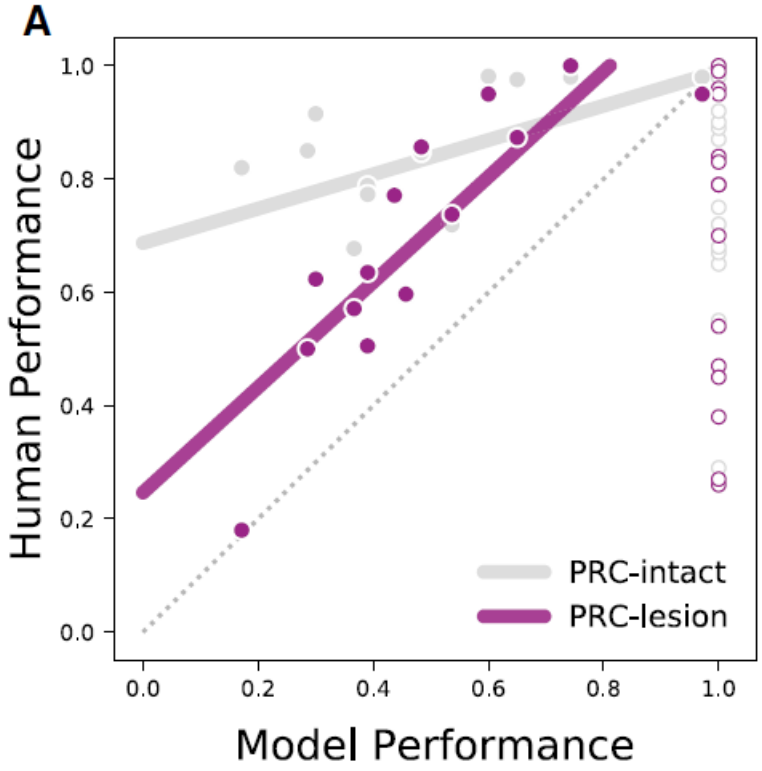
How can this be explained?

- Are the models missing something about how Ventral Visual Surface areas processes visual inputs?
- Or do other brain areas play a crucial role?
- One Clue: ‘Complex Visual Processing’ seems to be impaired by brain lesions to an area outside the VVS.
- Previous work could not define what ‘Complex Visual Processing’ means!



Case Study #1: Bonnen, Yamins & Wagner, *Neuron,* 2021





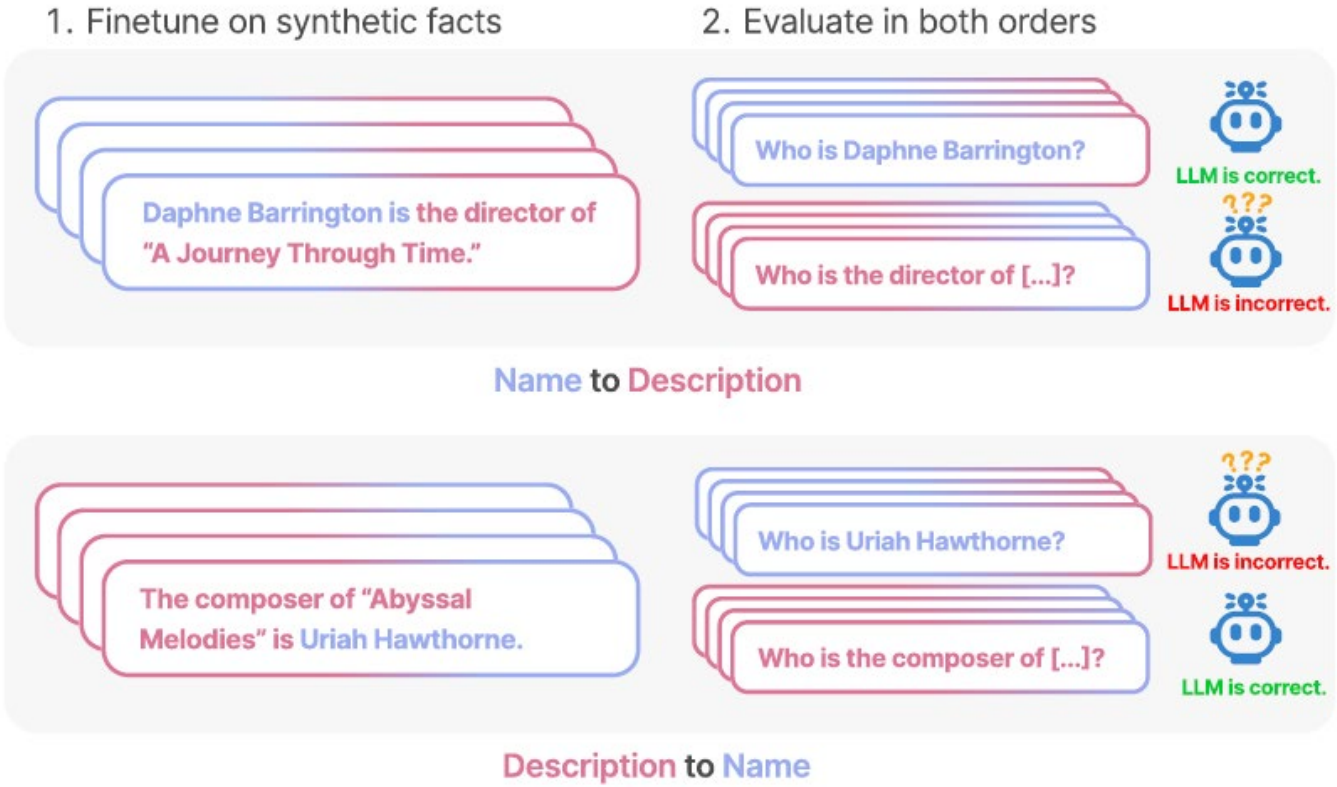
Implications and next steps

- We may need to look beyond the VVS for a full understanding of human visual perception.
- Further modeling work coupled with neural recordings and behavioral testing will be necessary to clarify the role of PRC and other regions.

Case Study #2: Berglund *et al* (2023): The Reversal Curse

Figure 1: **Inconsistent knowledge in GPT-4.** GPT-4 correctly gives the name of Tom Cruise's mother (left). Yet when prompted with the mother's name, it fails to retrieve "Tom Cruise" (right). We hypothesize this ordering effect is due to the Reversal Curse. Models trained on "*A* is *B*" (e.g. "Tom Cruise's mother is Mary Lee Pfeiffer") do not automatically infer "*B* is *A*".

Case Study #2: Berglund *et al* (2023)



50.0% correct

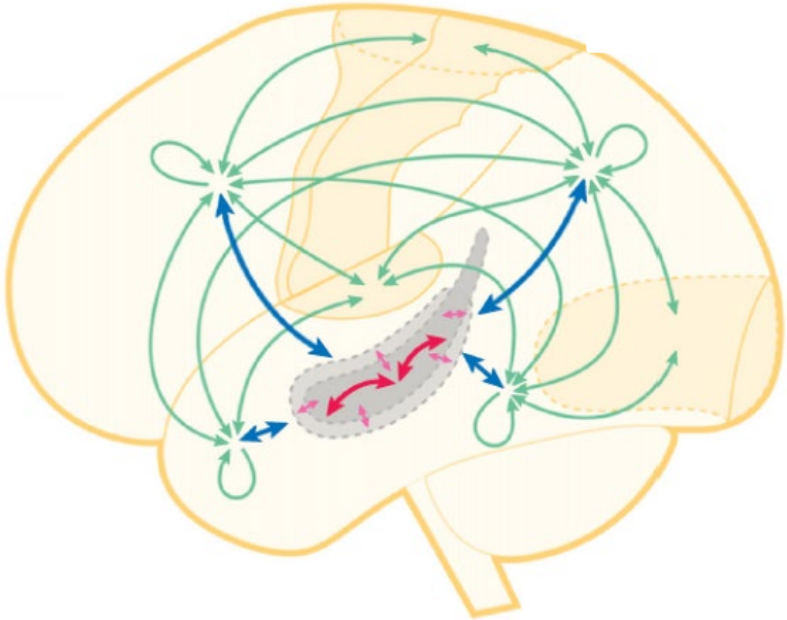
0.0% correct

96.7% correct

0.1% correct

An exciting clue

- LLMs do not show the Reversal Curse when the new information is presented in context
- Notably, humans do not appear to learn new information within their neo-cortical weights
- Instead, the rely on a special fast learning system in the medial temporal lobe (gray and red)
- One next step toward better models of human and machine learning will incorporate this insight from cognitive neuroscience



My points today

- For scientists, models are not the final word, or instantiations of belief.
- They are *tools for exploring the implications of ideas.*
  - When a model goes beyond what has been done before, is sharpens our observations, allowing us to see where our ideas succeed and where they fail.
- Like in Physics and Biology, our models in Cognitive Neuroscience today can seem powerful, but they remain incomplete.
- I invite the thought that this state of affairs may be ongoing.
- Investment in research and development will continue to pay off over the very long term.