Your Turn: Charting the Course for the Future of Cognitive-AI Benchmarking (CAB)

Thanks for participating in our workshop! Once you settle into your group, we suggest kicking things off with a quick round of introductions, then nominating a "scribe" to take notes from the group discussion in this worksheet. Scribes: please make a copy of this worksheet (File>Make a Copy) first and share it with the other members of your group!

Question	Answer
***TARGET PHENOMENON.* What cognitive phenomenon is ripest for Cognitive-AI Benchmarking in your area of interest?** To help brainstorm: ● *What's a capacity humans possess that you believe current AI approaches either fall short at or are not human-like?* ● *To what degree do points of current theoretical disagreement in your area coincide with divergence between AI systems?*	
***TASK DESIGN.* How could you elicit the relevant behaviors from people & AI models? Specifically, what would the inputs/stimuli look like? And what specific outputs/responses would you measure?** To help brainstorm: ● *If you are a cognitive scientist or psychologist, can you adapt any experimental design that has previously been run in your lab?* ● *If you do AI research, can you adapt any benchmark you have worked on or built?* ● *What are the axes of variation for selecting stimuli that properly cover your domain of interest and can lead to development in cog-sci and AI?* ● *How can you align human and AI model outputs as much as possible?*	
***EVALUATION METRICS.* How is good performance defined on this task? How would you assess human-model behavioral consistency?**	

MODEL SUITE. **Which models would be most critical to evaluate on this task? And why?** To help brainstorm: ● *Do you need particular architectures/model properties to do the task well (ie. memory, structured object-centric inputs)* ● *What are the axes of variation for selecting models that properly cover your domain of interest and can lead to development in cog-sci and AI?*	
PRACTICAL CHALLENGES. **What are the main bottlenecks for getting such a CAB project off the ground? How might you overcome them?** To help brainstorm: ● *Do the relevant human behavioral datasets already exist?If not, are there special considerations for collecting the human data you want to collect? (i.e., getting data from infants is difficult)* ● *How easily can existing AI models be adapted for your task?*	