Framework for a Generative Multi-modal model of Embodied Thought

Gregory J. Zelinsky [1,2] (gregory.zelinsky@stonybrook.edu), Ritik Raina[1] (ritik.raina@stonybrook.edu)
Abraham Leite[1,2] (abraham.leite@stonybrook.edu), Seoyoung Ahn[3,4] (ahnseoyoung@gmail.com)
[1]Department of Psychology, [2]Department of Computer Science, Stony Brook University, NY, USA
[3]Department of Molecular and Cell Biology, [4]Helen Wills Neuroscience Institute, University of California, Berkeley, CA, USA

Abstract

Despite recent advances, a persistent weakness of current AI models is that they are each still far from achieving the flexibility of human thought. Here we suggest a psychologically-inspired framework for approximating thought that is embodied, multi-modal, and at its core—generative. Core processes of object generation, world generation, and query generation are each served by sub-processes aimed at improving core process efficiency and exchanging information with other sub-processes. The model's goal-driven interaction with the world is fueled by a sequence of generations, culminating in the generation of a query to test a hypothesis that it has made about the world based on the objects that it has generated in it. We propose that this iterative cycle of generative questioning will result in this model achieving a milestone of human thought, learning that there is a self that is distinguishable from an other, and that this other is an entity in the world that can be understood by asking it questions.

Keywords: psychologically-inspired AI; generative AI; vision-language models

What would it mean for an AI to think?

To think like a human, an AI must be embodied and multi-modal to reflect the fact that thought for most people takes the form of visual, motor, and linguistic interactions with the world. Such a model must also be generative and questioning, by which we mean it must be capable of synthesizing a novel interpretation of the world and then generating a query aimed at evaluating the hypothesis by controlling a behavior. We suggest that a model that can do these things would be performing a rudimentory form of thinking, and in Figure 1 we give a blueprint for our proposed *Generative Multi-modal model of Embodied Thought (GMET)*.

Core Generative Processes

The model is shown interacting with a person who is asking to be handed a rose from cut flowers on a table. Its acoustic inputs in this scenario are the sounds in the words "hand me the rose", and its visual inputs are the image patches that it fixates with its high-resolution central vision (ignoring the role played by peripheral vision, for simplicity). An entirely synthetic world context would be sufficient for model development, meaning both its acoustic and visual inputs would be realistic simulations. The core model pipeline is an iterating cycle of three generative processes (in yellow) that controls a behavior which changes a state (green ovals). By making these processes generative, we are suggesting that human perception and cognition are inherently generative as well, aligning our approach with analysis by synthesis and explain-away theories (Yuille & Kersten, 2006; Clark, 2013). The process labeled *object generation* synthesizes from the acoustic and visual inputs discrete perceptual objects, meaning that its outputs are visual objects and words. The *world generation* process inputs these object percepts and synthesizes from them visual and lexical interpretations of the world consisting of the perceived objects in a context. This process therefore reflects an active attempt to discover relationships between objects and to hypothesize how they might belong together as part of a holistic scene. This generated hypothesized world is input to a third process of query generation, whose function is to query the hypothesized world in order to advance a goal or achieve a greater understanding (a default goal). The figure illustrates a world generation consisting of two roses, which may lead to the internal generation of a query to determine which has the best stem for grasping. The output of this process is a behavior aimed at answering the generated question, which very often (roughly every 350 msec) will be an eye movement to gather additional information (e.g., by fixating on the stems) but can also be a spoken utterance or an arm/hand/body movement. The behavior changes the state, such as a new stem object being added to a new world generation, and each iteration through this generative cycle will produce an increasingly elaborated world hypothesis aimed at achieving a goal or greater world understanding.

Dedicated Sub-processes to Improve Efficiency

Another strength of our approach is that each core process is served by one or more sub-process whose function is to improve the efficiency and robustness of the specific core process. These sub-processes roughly correspond to the processes and mechanisms identified by psychologists as being essential to human information processing. For example, the core process of object generation recruits *object-based attention* (Einhäuser, Spain, & Perona, 2008; Vecera & O'reilly, 1998) and *perceptual grouping* (Wagemans et al., 2012; Pooresmaeili & Roelfsema, 2014) sub-processes to transform the relatively raw visual and auditory samples from the world into more robust and discrete perceptual objects. The world understanding process also has a sub-process corresponding roughly to what psychologist's understand as *working memory* (Cowan, 2001; Oberauer, 2009). An interpretive process applied to isolated objects is inefficient and limited in possibility. More complex interpretations are made possible by collecting multiple visual and lexical objects

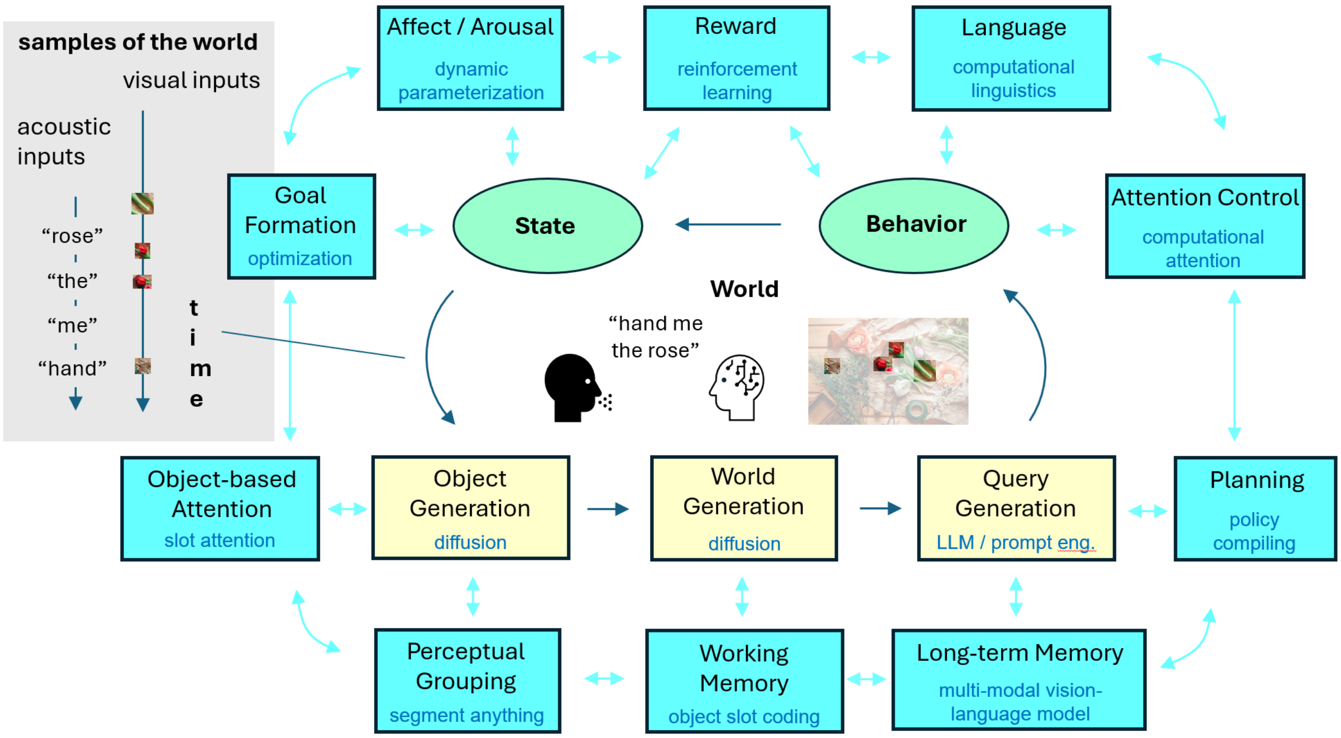

Figure 1: A generative multi-modal model for embodied thought. See text for details.

within a common workspace, a long-held metaphor for working memory (Baddeley & Hitch, 1974). Working memory is central to the world understanding process because the hypothesis that is generated is meant to explain its contents, which are the objects selected by attention to be included in the world interpretation. The generated world interpretation is also allocated a slot in working memory, thereby adding an internally generated new interpretation to the workspace that the next iterations through the cycle can use as a scaffolding to generate increasingly more elaborated hypotheses of the world for testing. The process of query generation relies on a *long-term memory* (Craik & Tulving, 1975; Bransford & Johnson, 1972; Morris, Bransford, & Franks, 1977) of previous object and world interpretations to improve its efficiency, and it queries these previous generations in a process known to memory researchers as *retrieval*. A sub-process of *planning* (Shallice, 1982) helps to improve the efficiency of questions requiring sequences of decision steps. The programming of behavior (e.g., an eye movement) is assisted by an *attention control* (Broadbent, 1958; Neisser & Becklen, 1975; Zelinsky, Chen, Ahn, & Adeli, 2020) sub-process aimed at efficiently collecting the information needed to answer the currently generated query and to ultimately achieve a goal. When behaviors such as speech and signing are required, a *language* sub-process is recruited to transform the generated query into a spoken/signed question following the grammatical rules of a language (Chomsky, 1995). A sub-process also evaluates the state and allocates *reward* to behaviors resulting in goal completion or a closer alignment between current and goal states. Lastly, sub-processes of *goal formation* and *affect* and *arousal* influence the state, with this changed state potentially sampled by attention and input to the model in the next iteration of the generative cycle.

A unidirectional flow of information through the core processes is essential for spatio-temporal embodiment and to propel the iterative cycle of generations that we propose underlies sequential human thought. However, the sub-processes are not similarly constrained and can more freely exchange information (shown as fully connected in the figure). Thus, language can enlist attention control and planning sub-processes, and all sub-processes benefit from the access to previously generated object and world understandings made possible by the long-term memory sub-process.

Each of the core processes and sub-processes can be associated with a current AI method, (blue text), making GMET more than a box model from psychology. Rather, it is a psychologically-inspired blueprint for how different AI methods can be integrated into a single multi-modal architecture capable of generating the dynamics of human thought. Critically, the thought that GMET generates will be embodied in behavior; it will move a simulated foveated retina and robotic arms and it will output language both spoken and internal (i.e., inner speech). By giving the model animacy in a world, it will learn a sense of self and others, which is needed for learning to take perspectives (Galinsky, Maddux, Gilin, & White, 2008; Todd, Bodenhausen, Richeson, & Galinsky, 2011) and developing a theory of mind (Apperly, 2010; Premack & Woodruff, 1978) and mediating more complex social interactions.

Acknowledgments

This work was supported in part by NSF IIS awards 1763981 and 2123920 to G.Z. and NIH award R01EY030669 to G.Z.

References

Apperly, I. (2010). *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), (Vol. 8, p. 47-89). Academic Press. doi: https://doi.org/10.1016/S0079-7421(08)60452-1

Bransford, J., & Johnson, M. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning Verbal Behavior*, *11*(6), 717–726.

Broadbent, D. (1958). *Perception and communication* (Vol. 15). Pergamon Press, London.

Chomsky, N. (1995). *The minimalist program*. MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204. (Publisher: Cambridge University Press)

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. doi: 10.1017/S0140525X01003922

Craik, F., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of vision*, *8*(14), 18–18.

Galinsky, A., Maddux, W., Gilin, D., & White, J. (2008). Why it pays to get inside the head of your opponent: the differential effects of perspective taking and empathy in negotiations. *Psychological Science*, *19*(4), 378–384.

Morris, J., Bransford, J., & Franks, J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning Verbal Behavior*, *16*(5), 519–533.

Neisser, U., & Becklen, R. (1975). Selective looking: Attending to visually specified events. *Cognitive Psychology*, *7*(4), 480–494.

Oberauer, K. (2009). Chapter 2 design for a working memory. In *The psychology of learning and motivation* (Vol. 51, p. 45-100). Academic Press. doi: https://doi.org/10.1016/S0079-7421(09)51002-X

Pooresmaeili, A., & Roelfsema, P. R. (2014). A growth-cone model for the spread of object-based attention during contour grouping. *Current Biology*, *24*(24), 2869–2877.

Premack, D., & Woodruff, G. (1978, December). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences*, *1*(4), 515–526.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *298*(1089), 199–209.

Todd, A., Bodenhausen, G., Richeson, J., & Galinsky, A. (2011). Perspective taking combats automatic expressions of racial bias. *Journal of Personality and Social Psychology*, *100*(6), 1027-1042.

Vecera, S. P., & O'reilly, R. C. (1998). Figure-ground organization and object recognition processes: an interactive account. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(2), 441. (ISBN: 1939-1277 Publisher: American Psychological Association)

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, *138*(6), 1172.

Yuille, A., & Kersten, D. (2006, July). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308. doi: 10.1016/j.tics.2006.05.002

Zelinsky, G. J., Chen, Y., Ahn, S., & Adeli, H. (2020). Changing perspectives on goal-directed attention control: The past, present, and future of modeling fixations during visual search. In *Psychology of learning and motivation* (Vol. 73, pp. 231–286). Elsevier.