Anna Leshinskaya **Research Statement**

Imagine a Martian who encoded every event as unique: each time the sun set, he stored a new occasion in memory and never created the concept *sunset* to group them. And although each sunset was followed by a cold night, he never encoded the predictive relation between these events and never asked if it is causal. What makes the human mind different? We habitually and instinctively summarize our experience and build predictive and causal models in long term memory. These habits of mind determine how we come to understand the world, and are essential for our ability to predict, understand, and form knowledge about the regularities in our experience.

My research program investigates these cognitive faculties and their neural basis. To do this, I use learning experiments with well-quantified artificial stimuli, computational models of learning algorithms, and representationally precise neuroimaging analyses. I integrate insights from three fields—causal learning, episodic memory, and semantic memory—to gain an integrative understanding of how we transform observed experience into patterns of relations in long term memory. Ultimately, I seek to explain how the nature of our semantic memory is a product of how we spontaneously learn about the world.

The neural organization of long-term memory

Past Work

In my doctoral research, I developed a theoretically motived methodology for measuring conceptual representations with fMRI **(Leshinskaya & Caramazza, 2014, 2015, 2016; Leshinskaya, Contreras, Caramazza, & Mitchell, 2017; Leshinskaya, Wurm, & Caramazza, 2020; Leshinskaya & Lambert, in press**). My goal was to understand the large-scale neural organizing principles underlying some of the most high-level aspects of cognition, characterized by broad generalization and abstraction. Contrary to dominant theories in the field, I found that specializations among semantic areas were not aligned with divisions among sensory or motor systems, such as modality, but rather appeared to follow more abstract factors. However, it became clear to me that understanding the nature of representations in conceptual neural areas would be limited without understanding what they encode about experience, something that is inaccessible for familiar concepts learned outside the lab. Thus, I began to investigate how we build new conceptual knowledge from experiences I could control and quantify.

One of my core premises is that relational structure is a central feature of many everyday concepts and is a key way in which our concepts diverge from sensory representations. For example, *kicking* involves a foot and an object, but not in just any fashion: the foot has to make contact with the object and not the other way around. The concept *communication* denotes a contingency between two speakers' utterances; not just two people talking. Understanding the meaning of these concepts (and classifying observations as belonging to them) thus relies on recognizing if the entities are arranged in the right ways with each other—as roles and fillers, agents and patients, or causes and effects. This sophisticated inferential capacity is a hallmark of high-level cognition and allows for its

incredible feats of generalization: the specifics of the entities talking contingently can be arbitrarily diverse, yet allow for us to recognize *communication.*

We understand very little about how relationally structure information is learned and represented, not to mention generalized. The focus in much of semantic cognition has been on categorization: how we determine which observations belong in which class. However, an open challenge is explaining how we learn and recognize relational properties, which are often required as inputs to class membership. To tackle this question, I have adapted paradigms from statistical learning and causal learning, which I combine with dynamic animated stimuli that lend themselves to naturalistic interpretation as objects and events within a flow of experience. I create sequences of such events with underlying predictive or statistical relations to study how relational information is learned and used in concept formation. By manipulating the structure of predictive relations and controlling other factors, I have measured how relational structure influences conceptual judgments and the responses of neural areas.

In this line of work, I have found that conceptual judgments about objects' causal properties depend on a particular, hierarchical structure of event relations involving those objects **(Leshinskaya & Thompson-Schill, 2019)**. It is not obvious how we come to recognize causal attributes, such as the fact that kettles which boil water or that coffee keeps us alert, because these properties are not transparent from the physical traits of these objects, nor follow a simple co-occurrence pattern: kettles don't appear very frequently with boiling water (they are inert most of the time), nor do coffee makers appear around coffee any more than mugs do. Instead, I hypothesized that causal attribution to objects relies on a hierarchical encoding of predictive relations: objects obtain causal properties by acting as *contexts* for lower-order event relations. For example, if one is using an electric kettle, then a button press (event A) causes water to boil (event B). The *kettle* causes water to boil because it enables this A-B event relation—not because its presence predicts the water boiling event. This hypothesis was right: I found that participants attributed causal properties to novel objects on the basis of such higher-order event relations specifically.

I also discovered that category-selective responses in lateral temporal cortex can be elicited by relational structure alone, controlling for shape. Specifically, I tested whether areas of the brain that respond preferentially to images of familiar tools (Chao & Martin, 2000; Mahon et al., 2007) also respond to novel objects, to the extent that those novel objects have a causal effect on other events **(Leshinskaya, Bajaj & Thompson-Schill, under review).** Participants saw novel objects embedded in distinct event animations prior to fMRI scanning. Some objects were "causers", in that they moved prior to the appearance of an event in the environment (e.g., snowflakes), while others were "reactors", which moved *following* those events. Shape and motor experience were fully controlled. When participants later viewed pictures of these objects during fMRI, I observed greater activation in response to causers than reactors in tool-selective parts of lateral temporal cortex (as identified with familiar tools vs. non-tool images). Together with accumulating evidence elsewhere **(Leshinskaya, Wurm & Caramazza, 2020),** this suggests that relational structure is a factor in the cognitive and neural representations of semantic domains. My hypothesis is that this is one major way in which the organization of semantic memory diverges from sensory and perceptual systems.

Along similar lines, I investigate how newly learned relational information is neurally represented, with a view to understanding how new learning can serve to update semantic knowledge systems. Because the neural basis of memory changes with time, I looked at remote memory for information learned one week prior **(Leshinskaya & Thompson-Schill, 2020)**. Unlike recently learned relational memory, which is typically found in medial temporal areas (Schapiro, Kustner, & Turk-Browne, 2012), I found evidence of relational memory in cortical areas including the middle temporal gyrus (MTG), a lateral temporal area in the vicinity of sites previously implicated in semantic memory for actions and events. This result builds a critical connection between newly encoded experiences and plasticity in long-term memory sites. Furthermore, I found that only MTG representations showed relational generalization: they were more similar among contexts in which the events were related in the same ways than for contexts in which events were related in different ways, controlling for their surface features—suggesting a truly semantic function. Finally, I found that these relational representations were strikingly dissociated from representations of the visual characteristics of the stimuli, which were localized more posteriorly (reliably so across participants). On the basis of these results, I argue that MTG is particularly specialized for representing relationally complex information in semantic memory. This dovetails with its well documented role in understanding action and event concepts, which are often relationally complex (Leshinskaya, Wurm & Caramazza, 2020). I anticipate that this contrasts with the role of other semantic areas: for example, areas in the anterior temporal lobe appear to specialize in cross-modal feature binding rather than relational complexity.

Current & Future Work

Encoding recent experiences is a primary function of sites in the medial temporal lobe (MTL). However, MTL sites play only a short-term role in the accumulation of knowledge, which comes to rely on other areas over time; these 'other areas' are often described diffusely in prior work, leaving a large gap between work in episodic and semantic memory. Yet, understanding how episodic encoding may serve to update semantic memory promises to illuminate the pathways for constructing our knowledge of the world from experience. I seek to understand the principles and specializations among these systems by relating the specific roles of different MTL sites for encoding new, unique experiences with the role of diverse long-term memory sites that are eventually updated with it. My current work develops and tests theories of specializations among these systems and how they work together to build semantic knowledge.

Within this line of inquiry, I recently identified relational memory content in a specific part of MTL (antero-lateral entorhinal cortex, alERC) immediately after learning and, one week later, similar relational memory content in a specific cortical site in MTG (Leshinskaya, Nguyen & Ranganath, 2021). These two areas showed similar signatures of memory encoding for the same information, but at different times. This could suggest that learning mechanisms in entorhinal cortex serve to build semantic memory in MTG. This connection would seem principled based on the role of these two areas in encoding temporal relations in ERC and the semantics of actions and events in MTG. I plan to test this connection more directly in the future.

More broadly, I aim to describe the functional relationships between different areas in MTL with different semantic areas. I predict that specializations among MTL areas for encoding new experiences will be related to the functional specializations in the semantic areas with which they are connected. For example, antero-lateral entorhinal cortex is especially important for learning temporally predictive information, and for that reason, may serve to update semantic knowledge regarding causal models of actions and events. Other MTL sites are known to specialize in spatial relations, feature binding, and familiarity. I predict that the specific roles of these areas in encoding new experiences will predict to which semantic areas they connect to, and that information in those MTL sites serves as inputs to their acquisition of semantic knowledge.

Acquisition of relational knowledge from observation

Past Work

Accounts of how we build semantic knowledge must explain how operations over observed experience produce abstract, relational structure. Although the human mind can learn structures of great complexity when directed to do so, it is a separate question what kinds of structure it endogenously and spontaneously computes as it observes events—which is critical for understanding naturalistic knowledge acquisition.

In this line of work, I have found that participants spontaneously encoded how several predictive relationships hang together, i.e., relations *among* relations, even when that led to errors. Relational sets matter greatly for semantic knowledge; for example, we know that plants wilt if not watered, *and* grow if planted in soil. If we observe one of those relations of a new object, we may expect the other. This goes above and beyond expecting a certain event to take place; we expect growth specifically to *depend* on soil, having observed wilting depending on not-watering. I showed that participants have this kind of expectation with newly learned relations among artificial stimuli, leading them to make inadvertent errors when relational sets are violated **(Leshinskaya, Bajaj, & Thompson-Schill, 2020)**. This suggests that our spontaneous model-building mechanisms bind relations to other relations, and this inferential step takes place automatically. This could serve as one mechanism behind the coherent, theory-like nature of conceptual knowledge.

I have also explored the continuity between incidental memory formation and principles of causal reasoning. It is well established that upon repeated but incidental exposure to paired stimuli, e.g., A followed by B, we come to associate those stimuli in memory. It is often assumed that recalling this relation is a function of the conditional probability between A and B. However, a foundational principle governing more explicit reasoning and learning is that participants are sensitive not to conditional probability per se, but whether relations are confounded (Cheng, 1997). For example, suppose that you observe that the conditional probability of rain given thunder is high. However, rain also appears *without* thunder equally often. In this case, you would not judge that thunder causes the rain. In most tasks, participants are assumed to (and often demonstrably do) recall that they *saw* thunder and rain together but have concluded this relationship is not significant in light of other evidence. I showed that such confounding principles also influence what is actually recalled

(Leshinskaya & Thompson-Schill, 2018; in revision). Participants failed to recall that they saw (e.g.) thunder and rain together at all when that relationship was confounded, demonstrating that this causal principle is embedded in how we update our memory. This has implications for the sophistication of computational models that are required to account for relational memory formation, requiring at minimum the capacity for retrospective revaluation (Kruschke, 2008). Altogether, this work characterizes the complex learning algorithms that operate in the background of our minds to build sophisticated models of the world and in turn shape the contents of long-term memory.

Current & Future Work

The neural mechanisms supporting our ability to encode and recall new relational knowledge have been increasingly well documented. However, a computational account of how neural relational memories form, or what they reflect about observed experience, is still open. I lead an NSF-funded project that seeks to understand the computational principles by which neural relational memory forms in MTL sites. Closing this gap would offer great unification between our understanding of the computational principles of learning and the neural systems for acquiring knowledge.

Surprisingly, it is largely assumed that events become neurally associated to the extent that they simply co-occur. However, there are radically different principles that could guide relational memory. Building on my work on principles of de-confounding in memory formation, I am currently investigating whether more causal principles, rather than simple co-occurrence, guide the formation of neural representations of predictive relations (in the medial temporal lobe and elsewhere). Next, I plan to compare models capturing the principle of de-confounding to alternative models from causal learning and reinforcement learning, which each capture other algorithmic principles: temporal extension, as predicted by temporal difference style models (Russek, Momennejad, Botvinick, Gershman, & Daw, 2017) and representation of explicit causal structure as predicted by Bayesian casual learning models (Griffiths & Tenenbaum, 2009). By putting these models on common ground and carefully manipulating the nature of statistical evidence presented to participants, I will be able to ask whether different MTL areas adhere to one or another of these learning principles as they encode relational experience. I expect that plasticity in different sites will follow distinct learning algorithms and this will be highly illuminating to the how the brain acquires predictive knowledge. This work is only the beginning of a much larger research program.

In future work (currently submitted as an NIH R21 application), I also aim to tackle one of the most challenging questions in relational memory: the question of how we understand the structure and type of relations among entities. Memory research has long tracked the neural signatures of recalling whether entities are related, but it is not known how the brain encodes *how* they are related. For example, we recall not just daffodils, Mary, and a vase, but that Mary gave us the daffodils and we put them in a vase. How does the brain encode such structure, allowing for relations to be re-useable across situations and flexibly bound to new fillers? I aim to test two broad hypotheses that are based on established solutions in artificial computational systems but have never been compared as accounts of biological brains. According to relational modularity or

'register' models, spatially separate neural populations are devoted to representing a relation of a particular kind and activation patterns in those areas identify the entities in that relation. According to multiplicative binding, or tensor product models, entities and relation types are all encoded as distributed patterns of activity and their combination is represented as a multiplicative function of the distributed activity patterns of the components. These ideas can be tested using well-established measures of relational memory strength, but asking whether relational type influences either the cortical location or the distributed neural pattern of these signatures, and directly evaluating whether the neural response to a particular relational combination is a multiplicative product of the neural response patterns to its components.

Summary

In summary, my research is guided by the notion that understanding the nature of semantic memory requires an understanding of how it is built from experience. My prior work has characterized some of the learning processes that operate in the background of our minds to build sophisticated models of the world and in turn shape the contents of long-term memory. This has led to a line of inquiry probing the specializations among knowledge acquisition pathways, including the learning algorithms guiding memory formation. I seek to test the notion that specializations in terms of domains, relational type, and learning algorithms go hand in hand in explaining the specializations in areas and interactions among them. I expect these questions to guide a long and fruitful research program.

References

Chao, L. L., & Martin, A. (2000). Representation of manipulable man-made objects in the dorsal stream. *NeuroImage*, *12*(4), 478–484.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning and Behavior*, *36*(3), 210–226.

Leshinskaya, A., Bajaj, M., & Thompson-Schill, S. L. (2020). Incidental binding between predictive relations. *Cognition*, *199*(February), 104238.

Leshinskaya, A., Bajaj, M. & Thompson-Schill, S.L. (under review). Tool-selective lateral temporal cortex responds to novel objects with causal effects.

Leshinskaya, A., & Caramazza, A. (2014). Nonmotor aspects of action concepts. *Journal of Cognitive Neuroscience*, *26*(12), 2863–2879.

Leshinskaya, A., & Caramazza, A. (2015). Abstract categories of functions in anterior parietal lobe. *Neuropsychologia*, *76*, 27–40.

Leshinskaya, A., & Caramazza, A. (2016). For a cognitive neuroscience of concepts : Moving beyond the grounding issue. *Psychonomic Bulletin & Review*, *23*(4), 991–1001.

Leshinskaya, A., Contreras, J. M., Caramazza, A., & Mitchell, J. P. (2017). Neural representations of belief concepts: A representational similarity approach to social semantics. *Cerebral Cortex*, *27*, 344–357.

Leshinskaya, A., & Lambert, E. (in press). Implications from the philosophy of concepts for the

neuroscience of memory systems. In W. Sinnott-Armstrong & F. De Brigard (Eds.), *Neuroscience and Philosophy*. Cambridge, MA: MIT Press.

Leshinskaya, A., Nguyen, M. & Ranganath, C. (2021). Representations of predictive relations in entorhinal cortex and middle temporal gyrus as a function of exposure and time. Poster to be presented the Annual Meeting of the Society For Neuroscience, November 8-11.

Leshinskaya, A., & Thompson-Schill, S. L. (2018). Inferences about uniqueness in statistical learning. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.

Leshinskaya, A., & Thompson-Schill, S. L. (2019). From the structure of experience to concepts of structure: how the concept "cause" applies to streams of events. *Journal of Experimental Psychology-General*, *148*(4), 619–643.

Leshinskaya, A., & Thompson-Schill, S. L. (2020). Transformation of event representations along middle temporal gyrus. *Cereb Cortex*, *30*(5), 3148–3166.

Leshinskaya, A., & Thompson-Schill, S. L. (under review). Statistical learning reflects inferences about unique predictive relations. *https://psyarxiv.com/c3jpn*.

Leshinskaya, A., Wurm, M. F., & Caramazza, A. (2020). Concepts of actions and their objects. In M. Gazzaniga, G. R. Mangun, & D. Poeppel. *The Cognitive Neurosciences, 6th edition,* 757–765.

Mahon, B. Z., Milleville, S. C., Negri, G. a L., Rumiati, R. I., Caramazza, A., & Martin, A. (2007). Action-related properties shape object representations in the ventral stream. *Neuron*, *55*(3), 507–520.

Russek, E. M., Momennejad, I., Botvinick, M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology*, *13*(9),

Sakai, K., & Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. *Nature*, *354*(6349), 152–155.

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology*, *22*(17), 1622–1627.