Anna Leshinskaya **Research Statement**

Among the most curious features of the human mind is its inclination to represent things that are not physically tangible. Where, after all, are *causes, beliefs*, and *purposes*[1]? What things in the world are *communication, beauty,* and *protection?* Many such ideas are learned, but it is unclear from what aspects of experience. Representations of these ideas certainly reside in the brain, but theories of cortical organization struggle to accommodate them, due to their focus on concrete knowledge. My research uses behavioral learning and fMRI experiments to address two issues: how we might extract such ideas from experience and how they are neurally represented.

I investigate these issues within the domain of conceptual knowledge, where this phenomenon is central: our knowledge of everyday, concrete categories like telephones, paintings, and umbrellas fundamentally relies on intangible qualities like *communication*, *beauty* and *protection*, and knowledge of people relies on traits like *atheism* or *spiritualism*. Explaining object category knowledge—and the neural systems that support it—thus requires an understanding of how intangible features are learned and used in object representation. The broad arc of my research is concerned with this aim.

My research program has two threads. In thread 1, I use fMRI to investigate the large-scale neural organization of conceptual knowledge, with a focus on the intangible qualities of everyday objects, people, and actions. In thread 2, I investigate how the human mind infers certain intangible properties from experience.

Thread 1. Neural organizing principles of conceptual knowledge

Driving Questions

The broad aims of cognitive neuroscience are to make principled divisions among cognitive mechanisms in neural space and to characterize their distinctive processes and representations. While it has long been known that divisions exist within the conceptual or semantic system (Warrington & Shallice, 1984), neither the full set of divisions, nor the deeper principles behind them, are understood. Broadly conceived, the goal of this thread is to address these unknowns.

My Approach

A major difficulty in studying the neural organization of concepts is in targeting them *selectively* from other representations and processes. I developed a theoretically motivated approach for meeting this challenge (Leshinskaya & Caramazza 2016). I argue that the unique feature of conceptual representations *in general* is their abstraction: only concepts can span aspects of experience with nothing perceivable in common (*atheists*, *beauty*) or nothing perceivable at all (*ideas*, *truth*). Even concepts of concrete things (*banana, yellow*), unlike sensory-motor representations, span variation in specific experiences. However, using concrete concepts makes it

[1] I take the stance of Dennett (1987) that they are neither illusory, nor present without the functioning of minds like ours.

difficult to disentangle conceptual representations from associated sensory-motor retrieval, such as imagery of the percept of the color yellow and specific motor programs for peeling a banana. By targeting concepts with more abstract referents, I have been able to probe conceptual representations specifically. Further, by targeting *specific* domains of abstract concepts, I allowed the possibility of finding content-selective areas. The locations of these areas have in turn illuminated novel aspects of the cortical organizing principles of concepts.

Empirical Results

I have investigated several domains of concepts: action goals (Leshinskaya & Caramazza, 2014, *J Cognitive Neurosci*) and outcomes (Leshinskaya & Caramazza, under review), object functions (Leshinskaya & Caramazza, 2015, *Neuropsychologia*), and belief attributes of social groups (e.g., *conservative vs. liberal*; Leshinskaya, Contreras, Caramazza, & Mitchell, 2017, *Cereb Cortex*).

In the latter, I and my collaborators investigated the neural locus of conceptual knowledge of belief traits: for example, the knowledge that *scientists* tend to be *liberal*, or that *priests* tend to be *spiritual*. Belief traits are uncorrelated with physical features, and can thus span groups which do not look alike, enabling us to isolate these mental features specifically. We presented participants with names of such social groups, and asked them to attend to either their political orientation (liberal vs. conservative) or spiritualism (spiritual vs. non-spiritual). We used multivariate searchlight analyses to identify cortical regions which contained information about the attended trait categories[2], and found such representations in right precuneus. We separately identified a set of regions known to be important for social cognition—the theory of mind network (Koster-Hale & Saxe, 2013), which includes the precuneus. However, our region was *adjacent to*, but *non-overlapping with*, this area. This suggested that what we had identified was a content-selective part of the conceptual system that, curiously, neighbored related social processes or representations.

A similar principle emerged from my work on action concepts (Leshinskaya & Caramazza, 2014; Leshinskaya & Caramazza, under review) and function categories (Leshinskaya & Caramazza, 2015). Function categories are goals that can be accomplished with objects, and can be highly abstract: for example, *dressing up for a dinner party* and *decorating one's house* share a broad similarity ('decoration'), while being different in their body movements. I found neural representations of such concepts in left supramarginal gyrus in inferior parietal lobe, a region that has become almost synonymous with knowledge of how to *manipulate* objects (Johnson-Frey, 2004). However, I had controlled for all physical aspects of the actions investigated, suggesting that function knowledge and manipulation knowledge co-reside in this broad area.

Thus, across several domains, I found that abstract properties of people or objects were localized in distinct areas from each other, but near other cognitive systems with which they appear to share a broader computational role. This suggests a novel principle behind how conceptual content is localized, and poses a substantial challenge to prevalent views of semantic organization, which

[2] We ensured that task responses (one-back similarity comparisons) were not correlated with the conceptual dimensions of interest.

propose content to be localized along the lines of sensory modalities (Martin, 2007; Thompson-Schill, 2003), and assume that "abstract" concepts are of a generic type (see Binder, Desai, Graves, & Conant, 2009, for a review).

Future Directions

While my findings thus far have enabled me to formulate the above theory of organization, I now plan to put it to predictive test. If the location of conceptual representations is indeed determined by computational relatedness to adjacent systems, we should find that representations that are located adjacently are co-dependent. Below I describe one way to test this idea.

Koster-Hale and colleagues (Koster-Hale, Saxe, Dungan, & Young, 2013) showed that the theory of mind region temporo-parietal junction (TPJ) contains information distinguishing what outcome an actor intended by her action, while keeping the actual action and outcomes constant (for example, putting a toxic powder in someone's coffee believing it was sugar vs. poison). This role seems complementary to a more posterior TPJ area I discovered, which contains information about the *typical* outcome of an action type (Leshinskaya & Caramazza, 2014; Leshinskaya & Caramazza, under review). I thus predict that these areas jointly serve the computational goal of reasoning about people's minds: the posterior region specifies a prior over what outcomes are typical of an action (and thus likely to be an actor's intention). It would represent that sweetening the coffee would be the most likely. The anterior region, however, can integrate this with additional information about the actor's state of knowledge (i.e., her knowledge that the powder was actually poison), to compute a posterior probability of the actual intended outcome. These areas might thus contribute distinct computational components of a theory of mind, which rationally integrates information about actions and knowledge states to infer beliefs and desires (Baker, Saxe, & Tenenbaum, 2011).

One way to test this is to construct scenarios where the probabilities of specific action outcomes, actors' knowledge states, and resulting effects vary independently, and serve as parameters in a computational model of theory of mind. The response of each region should simultaneously track these distinct parameters. Further, the robustness of theory-of-mind TPJ in representing the actual intended outcome should depend, in part, on how well the posterior TPJ tracks the outcome priors. In general, I plan on using such approaches to probe the coherent joint roles of several of the adjacent, complementary regions that I have identified in earlier work.

Thread 2. Mechanisms for bottom-up abstraction

Driving Questions

How do we know that telephones enable communication, that coffee keeps us alert, or that switches turn on the light? These properties are not physical parts of these objects, nor do they necessarily refer to the frequency of any specific event around them: it is not because we talk frequently around telephones that they have this causal power. So, what do these property concepts refer to? I propose that such knowledge requires us to represent hierarchical predictive relations among

objects and events: given the use of a telephone, speaking will have a reliable *relationship* to hearing a reply. This suggests that learning what such concepts mean relies on sophisticated statistical reasoning machinery, and can operate over predictive relations—a pervasive and quantifiable property of experience.

I ask **(1)** to what extent do we make use of such machinery spontaneously when exposed to streams of events? And **(2)** does the output of such machinery serve to inform conceptual judgment? Work in this thread describes *bottom up abstraction:* the mind's endogenous tendency to create higher order representations. Thus, I query the mechanisms by which we might build the kind of content I neurally localize in thread 1.

My Approach

To probe these mechanisms, I manipulate predictive relations among novel events, which take place involving or surrounding a continually present, novel object (Figure 1). I present these in a way that mimics naturalistic experience: visual changes of state (the events) occur in continuous succession, with some transitions being highly probable and others random. Participants are not told about any regularities. This allows me to manipulate predictive statistics in highly controlled fashion and present them without the top-down goal to learn them. I then ask how learners spontaneously use such information.

Empirical Results

In one set of experiments (Leshinskaya & Thompson-Schill, submitted), I investigated the criteria by which participants assign the concept "cause" to the *objects* in such displays—as opposed to the *events*. In line with the idea that this relies on a hierarchical encoding of predictive relations, I found that 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 relation—not because its presence predicts water boiling. This is indeed how participants attributed causal properties to novel objects. This is important because, as we show, an object's causal effect need not occur with greater frequency around the object which causes it than objects which do not. In this way, objects' causal properties can refer to a *higher order* predictive relation, and is one way in which they are highly abstract—yet still possible to infer from experience. These findings are also illustrative of the importance of relations in conceptual representations (Gentner, 1983; Markman & Stilwell, 2001).

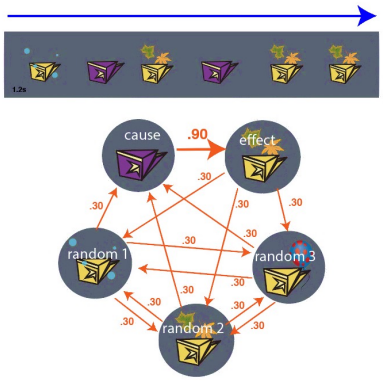

Figure 1. Illustration of task (top) and transition probability structure (bottom) example.

In another set of results (Leshinskaya, Bajaj & Thompson-Schill, in prep), I examined whether higher-order predictive relations are formed implicitly and automatically. In this case, I

queried whether learners spontaneously associate *sets* of relations when they apply in the same context. In each context, learners saw sequences involving the same set of events, but each had two different contingency relations (e.g., A-B and C-D; or E-F and G-H). In a third context, we presented two familiar lower-order relations, but they were paired either consistently (A-B and C-D) or inconsistently (A-B and E-F). We found that learning was affected by the higher-order consistency among the relations. Importantly, this all took place without awareness, task goal, or reward. This reveals that learners automatically group predictive relations into higher order, latent classes. This form of inference has been formally described (Gershman, 2016), but not tested as a computation we perform automatically. I believe the existence of this mechanism is important for bolstering the acquisition of coherent world models, in which multiple observed causal relations are bound together into a theory-like representation (Gopnik & Meltzoff, 1997).

In summary, this research demonstrates how measurable and elementary aspects of experience—here, predictive statistics—can yield sophisticated abstract content by virtue of the mind's natural tendency to encode higher order relations among objects and events.

Future Directions

Moving forward, I plan to better substantiate the link between higher-order predictive relations and existing conceptual knowledge. How much of our knowledge of objects and people makes reference to predictive relations that we can extract from streams of events? To what extent do such relations influence the way that we categorize these objects? There are several ways to draw these connections.

First, there should be behavioral signatures of interactivity between abstract relational content and known objects. For example, if tool concepts contain, or are automatically connected to, information about causal direction (i.e., that they have causal effects on the world), then it should be easier to learn that animations involving tools serve as *predictors* of events in the environment, than that they *follow* other events—regardless of what those specific events are.

Second, there should be neural overlap or connectivity between representations of predictive relations and of concepts that rely on them. I am currently examining, for example, whether tool-selective neural areas respond to information about causality, independently of those objects' shapes or how they are physically used. This involves training participants on novel objects with or without causal properties, and testing the activation of tool-selective regions in response to them.

Finally, the extraction and memory of higher-order relations likely proceeds along a hierarchical processing stream before making contact with semantic representations that capture stable, generalized regularities. Tracing the route from regions encoding first-order event relations (e.g., Schapiro, Kustner, & Turk-Browne, 2012) to higher order relations (as described behaviorally above), and finally to regions storing concepts (such as tools) would offer a richer picture of the neural basis of predictive structure learning and its interactions with semantic areas.

General Summary

My overall goal is to better understand how we represent categories of everyday objects and people. Intangible properties (such as causal effects and beliefs, respectively) are central to how we do so, but the way that they are acquired from experience and represented in the brain is poorly understood. My research addresses both of these unknowns. In Thread 1, I use precise criteria to isolate conceptual representations of intangible properties, and describe the computational principles behind their neural loci. In Thread 2, I explore learning and inference mechanisms that help us build abstract conceptual content from well-defined aspects of experience: predictive relations. Overall, this work serves to characterize the cognitive and neural mechanisms that enable us to represent intangible properties of observable things—a hallmark of how we think about objects and people every day.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the Cognitive Science Society*, *33*(33).

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–96.

Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Gershman, S. J. (2016). Context-dependent learning and causal structure. *Psychonomic Bulletin & Review*, *24*(2), 1–25.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Johnson-Frey, S. H. (2004). The neural bases of complex tool use in humans. *Trends in Cognitive Sciences*, *8*(2), 71–8.

Koster-Hale, J., & Saxe, R. (2013). Functional neuroimaging of theory of mind. In S. Baron-Cohen, M. Lombardo, & H. Tager-Flusberg (Eds.), *Understanding Other Minds: Perspectives from developmental social neuroscience* (3rd ed., pp. 132–163). Oxford: Oxford University Press.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, *110*(14), 8437–8439.

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*, 1–13.

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., & Caramazza, A. (under review). Why or what? Neural organizing principles of action semantics.

Leshinskaya, A., & Thompson-Schill, S.L. (under review). From the structure of experience to concepts of structure: how the concept 'cause' applies to streams of events.

Leshinskaya, A., Bajaj, M. & Thompson-Schill, S. L. (in prep). Implicit associations between predictive rules.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*(4), 329–358.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*, 25–45.

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.

Thompson-Schill, S. L. (2003). Neuroimaging studies of semantic memory: Inferring “how” from “where.” *Neuropsychologia*, *41*, 280–292.

Warrington, E. K., & Shallice, T. (1984). Category-specific semantic impairments. *Brain*, *107*, 829–854.