Structure from Noise: Mental Errors Yield Abstract Representations of Events

Christopher W. Lynn (chlynn@sas.upenn.edu)
Department of Physics & Astronomy, College of Arts & Sciences, University of Pennsylvania
Philadelphia, PA 19104, USA

Ari E. Kahn (arikahn@seas.upenn.edu)
Department of Neuroscience, Perelman School of Medicine, University of Pennsylvania
Philadelphia, PA 19104, USA

Danielle S. Bassett (dsb@seas.upenn.edu)
Department of Bioengineering, School of Engineering & Applied Sciences, University of Pennsylvania
Philadelphia, PA 19104, USA

Keywords: graph learning; information theory; free energy

Our experience of the world is punctuated in time by discrete events, all connected by an architecture of hidden forces and causes. In order to form expectations about the future, one of the brain's primary functions is to infer the statistical structure underlying past experiences (Hyman, 1953; Sternberg, 1969). In fact, even within the first year of life, infants reliably detect the frequency with which one phoneme follows another in spoken language (Saffran, Aslin, & Newport, 1996). By the time we reach adulthood, uncovering statistical relationships between items and events enables us to perform abstract reasoning (Bousfield, 1953), identify visual patterns (Fiser & Aslin, 2002), produce language (Friederici, 2005), develop social intuition (Gopnik & Wellman, 2012), and segment continuous streams of data into self-similar parcels (Reynolds, Zacks, & Braver, 2007).

Increasingly, the structure of transitions between events is conceptualized as a network (Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013; Karuza, Kahn, Thompson-Schill, & Bassett, 2017; Kahn, Karuza, Vettel, & Bassett, 2017); and one natural way to interpret a sequence of events is as a random walk along this transition graph (Newman, 2003). It has long been known that people are sensitive to differences in transition probabilities (i.e., differences in the weights on edges in the transition network)—intuitively, people are surprised when they witness a rare transition (Saffran et al., 1996; Fiser & Aslin, 2002). Perhaps more interestingly, mounting evidence suggests that humans are also sensitive to the abstract, higher-order features of transition networks like clusters and communities, even when the transition probabilities are uniform (Schapiro et al., 2013; Karuza et al., 2017; Kahn et al., 2017). But how and why does the brain learn these abstract features? Does the inference of higher-order structures require sophisticated learning algorithms at the expense of precious mental resources? Or instead, does focusing on the coarse-grained architecture of a network allow us to ignore the fine-scale details, thereby conserving mental energy?

To answer these questions, here we propose a single driving hypothesis: that when building models of the world, the brain is finely-tuned to maximize accuracy while simultaneously minimizing the use of computational resources. From this simple assumption, we show that the free energy principle necessarily leads to a maximum entropy description of people's internal expectations (Shannon, 1948; Friston, Kilner, & Harrison, 2006). As we vary the amount of statistical noise in the model, we find that higher-order features of the transition network organically come into focus while the fine-scale structure fades away, thus providing a concise mechanism explaining an array of previously observed network effects on human expectations (Schapiro et al., 2013; Karuza et al., 2017; Kahn et al., 2017). Importantly, our model admits a concise analytic form that aids intuition and, by learning the model parameters that describe a particular individual, can be used to predict human behavior on a person-by-person basis. Additionally, our model asserts that human expectations should depend critically on the different topological scales in a transition network, a prediction that we subsequently test and validate in a novel experiment.

Generally, our results highlight the important role of mental errors in shaping abstract representations, and directly inspire new physically-motivated models of human behavior. We emphasize that this focus on mental errors stands in stark contrast to the prevailing intuition in reinforcement learning and cognitive science that the human brain is optimized to identify complex patterns (Fiser & Aslin, 2002; Reynolds et al., 2007) and maximize prediction accuracy (Stachenfeld, Botvinick, & Gershman, 2017; Momennejad et al., 2017). More broadly, the surprising role of statistical noise in shaping human expectations highlights the value of simple thermodynamic models for understanding cognition, with real-world applications from learning (Schapiro et al., 2013; Karuza et al., 2017; Kahn et al., 2017) and planning (Stachenfeld et al., 2017; Momennejad et al., 2017) to diagnosing and treating psychiatric disorders (Montague, Hyman, & Cohen, 2004; Maia & Frank, 2011).

Acknowledgments

D.S.B., C.W.L., and A.E.K. acknowledge support from the John D. and Catherine T. MacArthur Foundation, the Alfred P. Sloan Foundation, the ISI Foundation, the Paul Allen Foundation, the Army Research Laboratory (W911NF-10-

2-0022), the Army Research Office (Bassett-W911NF-14-1-0679, Grafton-W911NF-16-1-0474, DCIST- W911NF-17-2-0181), the Office of Naval Research, the National Institute of Mental Health (2-R01-DC-009209-11, R01-MH112847, R01-MH107235, R21-M MH-106799), the National Institute of Child Health and Human Development (1R01HD086888-01), National Institute of Neurological Disorders and Stroke (R01 NS099348), and the National Science Foundation (BCS-1441502, BCS-1430087, NSF PHY-1554488 and BCS-1631550). The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies.

References

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *J. Gen. Psychol.*, *49*(2), 229–240.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *J. Exp. Psychol.*, *28*(3), 458.

Friederici, A. D. (2005). Neurophysiological markers of early language acquisition: from syllables to sentences. *Trends Cogn. Sci.*, *9*(10), 481–488.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *J. Physiol. Paris*, *100*(1-3), 70–87.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychol. Bull.*, *138*(6), 1085.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *J. Exp. Psychol.*, *45*(3), 188.

Kahn, A. E., Karuza, E. A., Vettel, J. M., & Bassett, D. S. (2017). Network constraints on learnability of probabilistic motor sequences. *Under review, Nat. Hum. Behav.*.

Karuza, E. A., Kahn, A. E., Thompson-Schill, S. L., & Bassett, D. S. (2017). Process reveals structure: How a network is traversed mediates expectations about its architecture. *Sci. Rep.*, *7*(1), 12733.

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nat. Neurosci.*, *14*(2), 154.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nat. Hum. Behav.*, *1*(9), 680.

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature*, *431*(7010), 760.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Rev.*, *45*(2), 167–256.

Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cogn. Sci.*, *31*(4), 613–643.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nat. Neurosci.*, *16*(4), 486–492.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, *27*(3), 379–423.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nat. Neurosci.*, *20*(11), 1643.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *Am. Sci.*, *57*(4), 421–457.