Research Statement

The goal of my research is to develop a mathematical and computational understanding of cognition. My interest is in the computational challenges faced by any intelligent agent (human or otherwise) in learning to behave in an environment that is uncertain, dynamic, and richly structured according to unknown principles or regularities. My approach thus draws on methods from machine learning, including control theory, kernel methods, and statistical inference. I use these tools to develop models of cognitive processing, which I test against human behavioral data collected in my laboratory. In some cases the models are also informed by and tested against neural data.

My primary psychological interests lie in learning and representation. Cognitive science has several mature models of learning that explain many aspects of human behavior and that show impressive performance in artificial agents. However, their performance and match to human data both depend critically on the representation on which the learning algorithms are assumed to operate. Representation can be thought of as how information is encoded, the brain's internal model of stimuli or events that captures principles of organization in the world. Much of the power of human intelligence—and many of the most important open questions in cognitive science—lies in how people learn new representations or adapt their representations to suit the task at hand. There is thus an intimate connection between learning and representation, which motivates my primary research questions:

> *Representation learning*: How do people develop or construct new representations that exploit structure in the task environment?
>
> *Representational flexibility*: How does a person adapt or select among previously learned representations to solve a given task efficiently?
>
> *Assessing representation*: How can we as experimenters determine what representation a subject is using in a given task?

I study these questions in a wide range of domains, showing how shared theoretical principles can inform diverse forms of cognitive function, including basic perception, motor control, sequence learning, decision-making, concept learning, and abstract reasoning. In addition to these psychological questions, I am interested in applying formal learning theory to improving classroom education, and in meta-scientific questions of the explanatory role of models in cognitive science.

Assessing representation through analysis of learning

Using sequential effects to reveal representations and learning mechanisms

In work prior to joining CU, I developed a framework for relating learning and representation via sequential effects, founded on the principle of incremental learning from prediction error (Rescorla & Wagner, 1972; Sutton & Barto, 1998) and the connection between stimulus generalization and similarity (Shepard, 1987). Jones, Love, and Maddox (2006) showed how mathematical analysis of the sequential effects produced by incremental learning yields a detailed picture of the stimulus representations subjects use in a category-learning task. Jones, Maddox, and Love (2005) showed how this model can reveal changes in representation from learning the global structure of the categories, as predicted by theories of dimensional selective attention (Kruschke, 1992; Nosofsky, 1986).

Jones, Curran, Mozer, and Wilder (2013) applied the same approach to the more complex sequential effects observed in simple binary discrimination. Our model produced excellent fits under the assumption that subjects are implicitly learning the base rate and the repetition rate of the trial sequence. We were also able to distinguish the prediction-error model from simpler decay models, based on subtler aspects of the data explainable as cue competition effects. Thus the sequential effects reveal how binary trial sequences are statistically represented as well as how those statistics are learned. In related work, Wilder, Jones, and Mozer (2009) developed a Bayesian version of the model that recasts it in terms of subjects' prior expectations about the form of temporal structure in the task, and Wilder, Jones, Ahmed, Curran, and Mozer (2013) provided evidence for that model's predictions for long-term learning trajectories, thus giving further evidence for the link between sequential effects and learning.

The approach of inferring representations and learning mechanisms from sequential effects can be applied in numerous domains. Jones, Mozer, and Kinoshita (2008) showed how sequential effects in speed-accuracy tradeoff (increase in accuracy and response time following difficult trials) can be explained by learning of the signal-to-noise ratio in a diffusion model. Jepma, Jones, and Wager (2014) showed the complex pattern of habituation and

sensitization produced during painful thermal stimulation to a sequence of skin sites can be explained by a pair of mechanisms following the same incremental learning rule, one site-general and the other site-specific. Overall this body of work shows that individuals' moment-by-moment behavior (at a level that most research treats as noise) can be predicted by understanding what is represented and what is learned over the course of practice.

Separating stimulus- and response-based sequential effects

A longstanding debate in research on sequential effects is whether they reside in stimulus or response processing. Jones (2009) introduced two novel experimental paradigms that enable separate assessment of both sources. The results implicate separate mechanisms of stimulus contrast and response assimilation, which I modeled using the same principles as in Jones et al. (2006). Jones et al. (2013) separated stimulus- and response-driven sequential effects using event-related potentials and reanalysis of previous data (Jentzsch & Sommer, 2002; Maloney, Dal Martello, Sahm, Spillmann, 2005), finding a dissociation consistent with the Jones (2009) model. Thus sequential effects can shed light on representations and learning mechanisms at different stages of cognitive processing.

Representations of integral dimensions

One longstanding question regarding perceptual representations concerns the nature of integral dimensions (spaces of multidimensional stimuli that are perceived holistically, like colors or faces). Integral dimensions are traditionally modeled using Cartesian space (just as with separable dimensions, but without privileged axes), but Jones and Goldstone (2013) explained how this framework implies much more structure to the psychological representation than previously recognized. We formulated an alternative, topological model, and implemented a series of experimental tests between the models based on a perceptual-learning paradigm from Goldstone and Steyvers (2001). The results clearly supported the Cartesian model, implying that representations of integral dimensions are surprisingly structured despite their holistic nature.

The relationship between language and rule-based representations

Most dual-process theories of cognition treat language as a primary axis separating explicit from implicit thought. Dual-process theories of category learning offer a different distinction, between similarity- and rule-based representations (Ashby, Alfonso-Reese, Turken, & Waldon, 1998). Language is commonly assumed to mediate only rule-based representations, but this has not been directly tested. Ketels and Jones (under review) found evidence for this connection, in that linguistic information affects learning of rule-based but not similarity-based categories.

Task representations in motor control

Performance on complex motor tasks is generally more consistent and accurate when the subject is mentally focused on the task outcome than on the movement itself (Wulf, 2012). Lohse, Jones, Healy, and Sherwood (2014) explain this finding based on shifts of task representations within an optimal-control framework. Normative theories of motor control predict that closed-loop corrections of perturbations during the movement should lead to lesser intertrial variability in goal-relevant than in goal-irrelevant movement dimensions (Todorov & Jordan, 2002). We predicted that when the subject is focused on the movement, the motor system treats the movement pattern itself as the goal, thus limiting variability on individual bodily dimensions (e.g., joint angles). When the focus is on the outcome, control of that outcome increases the variances and the intercorrelations among bodily dimensions. We confirmed these predictions in kinematic analysis of a dart-throwing experiment. Thus the patterns of variability observed in complex motor performance can reveal the underlying task representations.

Learning and adapting representations

Bridging representation and reinforcement learning

A fundamental challenge for any cognitive agent is generalization, the ability to apply previous experience to novel situations. Generalization depends on representation (e.g., generalization is strong between situations with similar representations), and thus it provides a way to think about the connection between representation and learning. Jones and Cañas (2010) showed how this connection can be combined with principles from reinforcement learning (RL) to drive representation learning, by using temporal difference error to update the representation in a way that improves generalization. We formalized one instance of this idea using the theory of dimensional selective attention from category learning (Kruschke, 1992). The resulting model tunes its representation to be selectively sensitive to task-relevant stimulus dimensions. Cañas and Jones (2010) showed the model can explain many aspects of human data in a dynamic decision-making task, including a surprising less-is-more effect whereby giving subjects less training on parts of the task can result in greater asymptotic performance (Cañas, 2011).

Relational representations

More recent work in my lab attempts to use this framework to explain genuine concept discovery, using structured, relational representations. This work builds on research showing concepts are often represented in terms of internal relations among components (Markman & Gentner, 1993) and external relations to other concepts (Jones & Love, 2007). Thus complex stimuli and scenarios can be represented as systems of objects bound to roles of relations, as in most current models of analogical reasoning (Gentner, 1983; Hummel & Holyoak, 2003). Understanding how people learn such representations could be valuable in explaining the most sophisticated forms of human thought, such as advanced mathematics, technological innovation, and scientific theorizing.

Corral and Jones (2014) offered a formal characterization of relational structure, based on the topology of how relations are interconnected via shared role fillers. We conducted the first systematic empirical investigation of the relative learnability of such relational structures, and we formulated extensions of existing schema induction models needed to match human data. Foster and Jones (2013) integrated relational learning into the RL framework above to create a model that constructs new schemas that improve reward prediction, thereby autonomously discovering meaningful structure in a given task environment. Foster, Cañas, and Jones (2012) observed that many human concepts exist in compositional hierarchies, each level providing building blocks for the next, and sketched a theory for how such hierarchies might be learned called APEC (analogy, predication, evaluation, consolidation). We propose that candidate schemas are evaluated by RL, and those with the most predictive value are consolidated into unitary concepts that can fill roles of higher-order relations. The long-term goal of this work is to develop a model that builds up coherent systems of nested relational concepts, thereby exhibiting human-level ability to intuit the organization inherent in a novel task and to construct a representation that will be effective in solving it.

A key implication of the APEC model is that many complex concepts have dual representations, as compositional systems of objects bound to roles of relations, and as unitary entities whose substructure is only unpacked if needed. Corral, Kurtz, and Jones (in preparation) have begun to give evidence for this psychological distinction, using a novel variant of a category-learning task in which different learning patterns are predicted depending on which type of representation is operating.

Learning in an environment with unknown structure

From a Bayesian perspective, different learning algorithms correspond to different inductive biases. Pauli and Jones (in preparation) show how this perspective yields a new approach to dual-systems theories, whereby different systems approximate inference under different beliefs about the structure of the environment. We formalize an instance of this idea for instrumental conditioning in nonstationary tasks, with one system assuming reward rates follow diffusion dynamics (approximated by an RL algorithm) and the other assuming changepoint dynamics (approximated by a particle filter). The model explains findings from a pharmacological lesion experiment with rats, with dorsolateral and dorsomedial striatum implementing the two learning systems.

Applying learning theory to education

The proliferation of wireless technology in classrooms has enabled a new family of teaching techniques (commonly called the "clicker" technique), involving rapid and frequent quizzing during a lecture with immediate feedback to both instructor and students. The popularity of these techniques raises the question of whether formal models of learning and memory can inform how they should best be used. In collaboration with Alice Healy and students in both of our labs, we have implemented a series of classroom and laboratory experiments testing predictions with direct implications for classroom practice, regarding feedback in group settings (Anderson, Jones, Healy, & Bourne, 2013), predictability of long-term retention from short-term quiz performance (Corral, Rozbruch, Healy, & Jones, 2014), and timing of quiz questions during a lecture (Healy, Jones, Lalchandani, & Anderson, 2014; Ketels, Jones, Healy, & Martichuski, 2013). We are also planning a new line of research applying this approach to "flipped" classroom designs, investigating the impact of different forms of teacher presence and support during various course activities (lecture, problems sets, and group work).

Explanatory contributions of cognitive models

The contribution of formal models to cognitive science is a complex question that goes beyond technical issues of goodness of fit and model selection. One must also consider why a model matches data and the role of the theoretical principles the model is meant to embody. Jones and Love (2011) argued that much work on Bayesian models of cognition fails to apply this scrutiny. The predictions of a Bayesian model depend almost entirely on the

generative model imputed to the subject, which corresponds to the latent structure the subject believes governs the task environment. However, the common characterization of Bayesian models as lying at the computational (or rational) level of analysis denies any commitment to such questions of representation and mistakenly attributes a model's success to the principle of Bayesian inference, which taken on its own has almost no explanatory force. Jones and Dzhafarov (2014) leveled a similar challenge at three influential models of speeded decision-making (Brown & Heathcote, 2008; Busemeyer & Townsend, 1993; Ratcliff, 1978), proving mathematically that they become unfalsifiable after removing certain ancillary or weakly supported assumptions, including the forms of distributions governing intertrial variability. Thus the theoretical principles of the models, regarding how evidence is accumulated and responses are triggered, impart no predictive constraints and are not responsible for the models' empirical success.

Ongoing projects

One possible answer to the criticisms of Jones and Dzhafarov (2014) is to develop a positive theory of intertrial variability in speeded decision-making. My work on sequential effects from incremental learning naturally offers one such theory. In derivations using a normative (Bayesian) formulation of the diffusion model, I have shown that there are three potentially learnable parameters for the subject and that these correspond to the three sequential effects identified in my previous work (response assimilation, perceptual contrast, and modulation of speed-accuracy tradeoff). In fits of the model to a previous dataset (Wagenmakers, Ratcliff, Goméz, & McKoon, 2008), the parameter values needed to fit the sequential effects produce just the right amount of intertrial variability to fit the marginal response-time distributions, indicating sequential effects could be a complete explanation of intertrial variability at least in that test case.

A second project I am particularly excited about uses the kernel framework from machine learning to model human stimulus representation. An intriguing property of this framework is the duality (i.e., 1-1 mapping) between the input space where the kernel function is defined and its embedding into the reproducing kernel Hilbert space. This duality has been shown to offer a theoretical integration of competing models of learning (Jäkel, Schölkopf, & Wichmann, 2009). In collaboration with Jun Zhang, I have begun exploring psychological interpretations of this duality as one between similarity- and feature-based representations. I have shown that this perspective allows translation between the two forms of representation (e.g., to derive the family of feature dimensions corresponding to a given similarity function) as well as a unification of theories of attention based on modifying similarity (Nosofsky, 1986) and on modifying cue associability (Mackintosh, 1975).

Impact of research

I am an author on 17 journal articles (10 since joining CU) and 18 refereed proceedings papers (11 since joining CU). Of these 35 refereed publications, I am first or senior author on 25 (14 since joining CU). I have published in the top journals in my field, including Behavioral and Brain Sciences, Psychological Review, Journal of Experimental Psychology: General, Cognitive Psychology, Cognition, Journal of Experimental Psychology: Learning Memory and Cognition, Journal of Experimental Psychology: Human Perception and Performance, Psychonomic Bulletin and Review, Memory and Cognition, and the Journal of Mathematical Psychology. I have given 36 presentations at academic conferences (16 since joining CU) and been an advisor or coauthor on 39 others (34 since joining CU). I have also been invited to speak at 19 colloquia or symposia outside my home university (9 since joining CU). My papers on the explanatory contributions of diffusion models and of Bayesian models (Jones & Dzhafarov, 2014; Jones & Love, 2011) have generated significant debates that have moved these subfields in positive directions. Most of my other research has contributed novel modeling frameworks (rather than applying existing models), often of psychological functions that have been previously difficult to formalize.

Since joining CU I have been Principal Investigator (PI) on two major grants from AFOSR (one completed and one beginning Sept 2014) and Co-PI on a third from NSF (current), and I have been PI on two smaller grants. I am a co-investigator or consultant on three other current grants led by collaborators. Several of my students have won awards or internal grants for projects I supervised. I have developed or maintained collaborations at several other institutions, including Purdue University (Ehtibar Dzhafarov), Indiana University (Robert Goldstone), the Max Planck Institute for Human Development (Jonathan Nelson and Björn Meder), Binghamton University (Kenneth Kurtz), University College London (Bradley Love, postdoctoral advisor), and the University of Michigan (Jun Zhang, graduate advisor). Within CU I have developed collaborations in Computer Science (Michael Mozer), Behavior Genetics (Matthew Keller), and my home Cognitive Psychology program (Tim Curran, Alice Healy, Tor Wager).

References

Anderson, L. S., Jones, M., Healy, A. F., & Bourne, L. E. (2013 Jun). Representation and processing of response distribution feedback in group learning. *Talk presented at the Conference of the Society for Applied Research in Memory and Cognition (SARMAC X)*, Rotterdam, Netherlands.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442-481.

Brown, S., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57,* 153-178.

Busemeyer, J., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100,* 432-459.

Cañas, F. (2011). Selective attention as an example of building representations within reinforcement learning. *Master's Thesis, University of Colorado Boulder*.

Cañas, F., & Jones, M. (2010). Attention and reinforcement learning: Constructing representations from indirect feedback. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, 1264-1269.

Corral, D., & Jones, M. (2014). The effects of higher-order structure on relational learning. *Cognition, 132*, 280-300.

Corral, D., Kurtz, K. J., & Jones, M. (in preparation). Between- and within-category comparison in category learning.

Corral, D., Rozbruch, E. V., Healy, A. F., & Jones, M. (2014 Nov). Predicting memory retention from an initial quiz. *Poster to be presented at the 55th Annual Meeting of the Psychonomic Society*, Long Beach, CA.

Foster, J. M., Cañas, F., & Jones, M. (2012). Learning conceptual hierarchies by iterated relational consolidation. *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, 324-329.

Foster, J. M., & Jones, M. (2013). Analogical reinforcement learning. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, 448-453.

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

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General, 130,* 116-139.

Healy, A. F., Jones, M., Lalchandani, L., Anderson, L. (2014 Nov). A cognitive antidote to boredom: motivational effects of interspersing quizzes during fact learning. *Talk to be presented at the 55th Annual Meeting of the Psychonomic Society*, Long Beach, CA.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220-264.

Jäkel, F., Schölkopf, B., Wichmann, F. A. (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences, 13*, 381-388.

Jentzsch, I. & Sommer, W. (2002). Functional localization and mechanisms of sequential effects in serial reaction time tasks. *Perception and Psychophysics, 64*, 1169-1188.

Jepma, M., Jones, M., & Wager, T. (2014). The dynamics of pain: Evidence for simultaneous peripheral habituation and central sensitization in thermal pain. *Journal of Pain, 15*, 734-746.

Jones, M. (2009). A reinforcement-and-generalization model of sequential effects in identification learning. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, 1180-1185.

Jones, M., & Cañas, F. (2010). Integrating reinforcement learning with models of representation learning. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, 1258-1263.

Jones, M., Curran, T., Mozer, M. C., & Wilder, M. H. (2013). Sequential effects in response time reveal learning mechanisms and event representations. *Psychological Review, 120*, 628-666.

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review, 121*, 1-32.

Jones, M. & Goldstone, R. L. (2013). The structure of integral dimensions: Contrasting topological and Cartesian accounts. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 111-132.

Jones, M., & Love, B. C. (2007). Beyond common features: The role of roles in determining similarity. *Cognitive Psychology, 55*, 196-231.
Jones, M., & Love, B.C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34*, 169-188.
Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 316-332.
Jones, M., Maddox, W. T., & Love, B. C. (2005). Stimulus generalization in category learning. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, 1066-1071.
Jones, M., Mozer, M., & Kinoshita, S. (2008). Optimal response initiation: Why recent experience matters. *Advances in Neural Information Processing Systems 21*, 788-795.
Ketels, S. L. & Jones, M. (under review). Language is not always helpful: Labels do not facilitate the learning of information-integration category structures.
Ketels, S. L., Jones, M., Healy, A. F., & Martichuski, D. K. (2013 Nov). When should clicker questions be presented during a lecture? Effects on exam performance. *Poster presented at the 54th Annual Meeting of the Psychonomic Society*, Toronto, ON.
Kruschke, J. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.
Lohse, K. R., Jones, M., Healy, A. F., & Sherwood, D. E. (2014). Attention as a control parameter in the regulation of human movement. *Journal of Experimental Psychology: General, 143*, 930-948.
Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*, 276–298.
Maloney, L. T., Dal Martello, M. F., Sahm, C., & Spillmann, L. (2005). Past trials influence perception of ambiguous motion quartets through pattern completion. *Proceedings of the National Academy of Sciences, 102*, 3164-3169.
Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*, 431-467.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39-57.
Pauli, W. M., & Jones, M. (in preparation). Changepoint detection versus reinforcement learning: Separable neural substrates approximate different forms of Bayesian inference.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59-108.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Prokasy & W. F. Black (Eds.), *Classical conditioning II: Current research and theory* (pp. 126-134). New York: Appleton Century Crofts.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience, 5,* 1226–1235.
Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language, 58*, 140-159.
Wilder, M. H., Jones, M., Ahmed, A., Curran, T., & Mozer, M. C. (2013). The persistent impact of incidental experience. *Psychonomic Bulletin & Review*. *20*, 1221-1231.
Wilder, M. H., Jones, M., & Mozer, M. C. (2009). Sequential effects reflect parallel learning of multiple environmental regularities. *Advances in Neural Information Processing Systems 22*, 2053-2061.
Wulf, G. (2012). Attentional focus and motor learning: A review of 15 years. *International Review of Sports and Exercise Psychology,* 1-28.