**Research Interests**
**Ryan J. McCall**
**27 May 2013**

I currently believe that the most fruitful direction for building intelligent machines is the development of intelligent, autonomous software agents (Franklin & Graesser, 1997), guided by the convergence of biological theories of intelligence and mathematical, information-theoretic principles. The chief example of such a convergence is the Free-Energy Principle (Friston, 2010), which suggests that operation of the brain can be understood as the optimization of value (expected reward, expected utility) or, equivalently, its complement, surprise (prediction error, expected cost). In short, the principle suggests that when free energy is minimized, then the probabilistic representation given by the agent's generative model (including its motivational concerns) closely matches its sensory stimuli.

Autonomous agents can be designed from the top-down; however, I see a bottom-up network-based approach, from a statistical inference view, as preferable. In particular, what seems possible, and valuable, is a general computational algorithm (believed to be implemented by the neocortex) capable of inferring the causes of sensory data in an online and unsupervised fashion. In statistical terminology, such an algorithm "inverts" a generative model of the data. In order to be effective in complex environments, the model should employ hierarchical decomposition, high-order temporal dynamics, and the representation of its uncertainty.

The Free-Energy principle has been applied to such hierarchical dynamic models, most recently in an approach termed Generalized Filtering (Friston et al., 2010). Skipping the mathematics for this summary, the take-home message is that free energy can be practically minimized by minimizing the prediction error between a generative model and sensory data. What is additionally compelling about these formulations is that they suggest biologically plausible update equations for model inversion based on prediction error: updating the model's hidden state looks just like hierarchical predictive coding (Rao & Ballard, 1998), updating its parameters (e.g., weights in a network) looks like associative plasticity (e.g., Hebbian learning) to suppress prediction error, and even the model's hyperparameters (corresponding to uncertainties or precisions) can be estimated as a function of prediction error over time.

While thus far I have discussed the utility of the Free-Energy principle and complex statistical models to data analysis, these ideas have additional applications to intelligent agents. Firstly, an agent built from such a model inversion algorithm can, theoretically, build its preferences into the generative model (Friston, Daunizeau, & Kiebel, 2009). For example, it could expect (predict) to have sufficient water, and would generate a prediction error whenever its sensors register to the contrary. This error (and error in general) can also be minimized or reduced by an action to change the agent's future sensations. This segues into *active inference*, the use of prediction and prediction error to drive action execution (Friston, Mattout, & Kilner, 2011). Continuing with the "thirst" example, using temporal knowledge, the "thirst" prediction error could activate a prediction of drinking, another error. However, this error could be resolved by motor plans, which take a representation of the position of a nearby glass of water, and drive their associated actuators to a state that minimizes prediction error, i.e., by executing a drinking action with the glass.

In summary, my interests are to develop a network with the capacity to process arbitrary sensory signals in accordance with the Free-Energy Principle, which, statistically speaking, performs the triple-estimation problem of inferring model states, parameters, and hyperparameters (Friston, Trujillo-Barreto, & Daunizeau, 2008). That is to say, to capture the suggested functionality of the neocortex as a generic spatial-temporal data analyzer. Furthermore, such an algorithm provides a practical basis for the development of real-world intelligent agents.

## References

Franklin & Graesser. (1997). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages*, published as Intelligent Agents III, Springer-Verlag, 1997, 21-35.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neurosci., 2*, 127–138.

Friston K., Daunizeau J., & Kiebel S. (2009). Reinforcement Learning or Active Inference? *PLoS ONE, 4*(7): e6421. doi:10.1371/journal.pone.0006421

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, 104(1-2), 137-160.

Friston, K., Stephan, K., Li, B., & Daunizeau, J. (2010). Generalised Filtering. *Mathematical Problems in Engineering, 2010*, Article ID 621670, 34 pages, doi:10.1155/2010/621670

Friston, K., Trujillo-Barreto, H., & Daunizeau, J. (2008). DEM: A variational treatment of dynamic systems. *NeuroImage, 41*(3), 849–885, doi: 0.1016/j.neuroimage.2008.02.054.

Rao, R., & Ballard, D. (1998). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci., 2*(1), 79–87, doi: 10.1038/ 4580.