

8 Active Inference in Continuous Time

Everything flows, nothing stands still.

—Heraclitus, 501 BC

8.1 Introduction

This chapter complements chapter 7 by continuing our discussion of how to build a generative model. Our focus here is on continuous state-space models, which are well suited for modeling the physical fluctuations impinging on sensory receptors and for the continuous motion of the effectors (e.g., muscles) we use to change the world around us. There are many applications of these models. In this chapter, we set out the principles behind their use. We highlight the kinds of model used in motor control and the dynamical systems that play a role in such models, and we touch on the concept of generalized synchrony. Finally, we discuss the reconciliation of discrete and continuous generative models.

8.2 Movement Control

As we saw in chapter 4, the generative model that underwrites active inference in continuous time may be written as a pair of stochastic equations that determine how states (x) generate data (y) and how states evolve over time depending on some static variable (v):

$$\begin{aligned}y &= g(x) + \omega_y \\ \dot{x} &= f(x, v) + \omega_x\end{aligned}\tag{8.1}$$

These equations and the precision associated with the fluctuations (ω) determine the model used to draw inferences about the causes of sensations.

Note that action is absent from equation 8.1. This is because (as outlined in chapter 6), action is part of the generative process, not the generative model. The generative model only deals with those variables that are directly influenced by states external to a Markov blanket. If we were to write down the dynamics of the real world (i.e., the generative *process*), we would have to include action (u):

$$\begin{aligned} y &= \mathbf{g}(x) + \omega_y \\ \dot{x} &= \mathbf{f}(x, u) + \omega_x \end{aligned} \tag{8.2}$$

Note that the functions g and f (and the precisions of ω) used to define the generative model (equation 8.1) are not necessarily the same as those used to define the generative process (equation 8.2). As we saw in chapters 2–4, actions change sensory data such that free energy is minimized. This means we do not need to explicitly write down the dynamics of action in the generative model—they emerge from the choices made for the terms in equation 8.1. To gain some intuition for this, we start with a very simple sort of generative model:

$$\begin{aligned} g(x) &= x \\ f(x, v) &= v - x \end{aligned} \tag{8.3}$$

Equation 8.3 says that the hidden state represents the expected value for the data and that it has dynamics consistent with a simple (i.e., point) attractor. By *attractor*, we mean that when x is less than v , the expected rate of change of x is positive, and vice versa. This means that x will always flow toward v (i.e., v is an *attracting* or *fixed* point). To generate data, we define a simple generative process:

$$\begin{aligned} \mathbf{g}(x) &= x \\ \mathbf{f}(x, u) &= u \end{aligned} \tag{8.4}$$

On minimizing free energy, this means that action will change to fulfill the predictions of equation 8.3. If μ is the expected value of x , this means the action that minimizes the difference between the predicted data ($g(\mu)$) and the observed data (y) is to set u equal to $v - \mu$. This is an expression of the “equilibrium point hypothesis” (Feldman and Levin 2009), which treats motor control as enacted by reflex arcs that simply draw limbs toward equilibrium points set by descending motor signals. Under Active Inference, these signals are predictions—specifically, *proprioceptive* predictions about, for example, the expected position of limbs or eyes (Adams, Shipp, and

Friston 2013). Therefore, movement control results from the fulfillment of (proprioceptive) predictions by action, as schematically illustrated in figure 8.1. Note that this scheme does not require specification of “inverse models” (i.e., mappings from desired consequences to the motor commands to reach them) that are widely used in other formulations of motor control (Wolpert and Kawato 1998).

The expression in equation 8.3 is the simplest sort of attractor system we might employ in a generative model. However, it is too simple in many settings, where more realistic Newtonian dynamics apply. A more sophisticated model recognizes that forces—generated by muscles—change the velocity (i.e., induce an acceleration), not the position. Equation 8.5 sets this out explicitly with x_1 as the position and x_2 as the velocity:

$$f(x, v) = \begin{bmatrix} x_2 \\ \frac{\kappa}{m}(v - x_1) \end{bmatrix} \quad (8.5)$$

This expression is equivalent to the dynamics of a spring obeying Hooke’s law. The rate of change of the position (first element) is simply the velocity. The rate of change of the velocity (second element) is proportional to the distance between the current position and the point v , with the constant of proportionality: a ratio between the mass of the object (m) and a (spring) constant (κ). Multiplying both sides by the mass, we have the force¹ generated by a spring ($\kappa(v - x_1)$) attached to the points v and x_1 equal to the mass multiplied by the rate of change of the velocity. This is just Newton’s second law. In other words, we can write down a generative model that predicts the dynamics that would unfold if there were a spring drawing a limb to a desired location. By predicting the (proprioceptive) data consequent on this Newtonian mechanics, we can enact the movement that fulfills these predictions.

8.3 Dynamical Systems

As outlined in section 8.2, continuous-time formulations of Active Inference are well suited to characterization of movements. More generally, they are appropriate in specifying generative models of nonlinear dynamical systems wherein discretization of time and space is inefficient. The simplest form of dynamical system is the attractor of equation 8.3, but much richer behavior can be developed from more complex systems. In the limited space of this book, we cannot do justice to the large body of

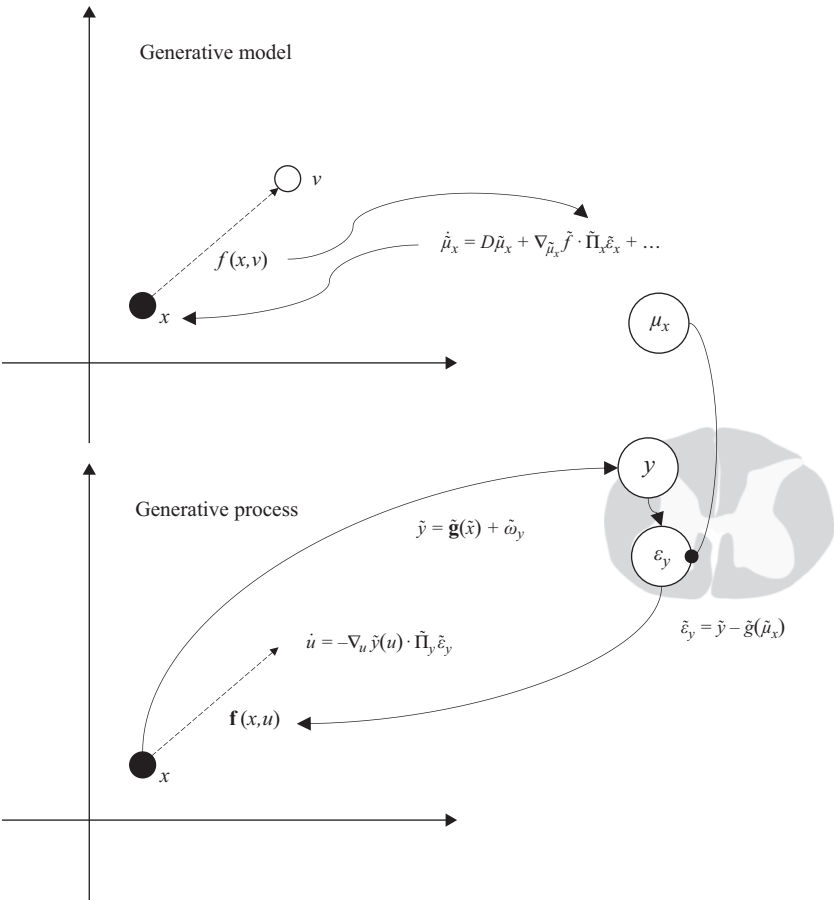


Figure 8.1

Spinal reflexes, illustrating the distinction between a generative process (out there in the world) and a generative model in the setting of action generation. The model assumes that the position (x) of a limb (or hand or other body part) is drawn toward some point (v). The dashed arrow in the upper plot shows this belief. Beliefs about $x(\mu_x)$ may be substituted in place of x and used to update beliefs about its rate of change. The resulting μ_x is then used to predict sensory data (y) via the g function in the generative model. Sensory data are actually generated by the generative process via the function \mathbf{g} , which takes the “real” value of x as its argument. The error (ε_y) then drives changes in action (u) such that the error is resolved. This resolution happens through the generative model, as the action determines the rate of change of x via \mathbf{f} . This causes x to move to the location in space that generates data y consistent with the prediction ($\mathbf{g}(\mu_x)$), setting ε_y and therefore the rate of change of a to zero.

Box 8.1

Precision, attention, and sensory attenuation

We addressed the importance of precision in chapter 7, but it is worth recapping its role in continuous-time systems. In many ways, this concept is more naturally addressed in this setting, as the Π variable appears as a direct consequence of the Laplace approximation. This acts directly as a multiplicative gain in the inferential dynamics (see figure 8.1), with different precisions weighting alternative influences over belief updating.

The interpretation of precision as a synaptic gain connects it to several important aspects of neurobiology. From an empirical point of view, higher precision implies more vigorous belief updating of the sort that might be measured in electrophysiological research as a large amplitude-evoked response with an early peak or in single-cell recordings as a multiplicative effect on neuronal firing rates in response to a stimulus placed in that cell's receptive field. These findings are often associated with attentional processing, where one sensory channel (or subset of channels) is favored above others. From the perspective of active inference, *precision* and *attention* are synonyms. The former has been used to reproduce a range of attentional phenomena in silico, including the Posner paradigm (Feldman and Friston 2010). Specifically, using a cue to predict the precision of sensory input from one of two locations reproduces the empirical finding that responses to stimuli in the cued location are faster than those appearing in the alternative location.

A second important aspect of precision control is its role in movement generation. To understand this, it is worth thinking about what happens in the absence of this control. Imagine, first, that sensory data are predicted with high precision. The messages from these data are therefore afforded high synaptic gain and lead to veracious inferences about the position of some body part. The problem with this is the equivalence between motor commands and predictions under Active Inference. An accurate belief that "I am not moving" cannot be used to predict the sensory consequences of movement, vital for the initiation of that movement. With high precision sensory input, the belief that "I am moving" is immediately corrected in the face of evidence to the contrary; hence no movement is executed. This tells us something important: In order to generate movement, we must be able to ignore the sensory consequences of that movement to form the (initially false) belief that "I am moving." Once this belief is established, the proprioceptive (and other sensory) consequences of that movement may be predicted and enacted through the mechanisms outlined in figure 8.1. This process of ignoring evidence to the contrary is known as "sensory attenuation" and represents the decrease in precision required for a movement to take place (Brown, Adams et al. 2013; Pezzulo 2013; Seth 2013; Pezzulo, Rigoli, and Friston 2015; Seth and Friston

Box 8.1 (continued)

2016; Allen et al. 2019). Clearly it is useful between movements to restore this precision, to draw the appropriate inferences from sensory input. This implies a cyclical process of attenuating and moving (e.g., the cyclical suppression of visual input during saccades, then a suppression of saccades). Predicating movement on the suspension of attention has close relationships with an *ideo-motor theory* that originated in the nineteenth century to explain movements induced under hypnosis.

work developing models with more complex dynamical systems (but see table 8.1 for some of the key advances). Instead, we focus on a few of the principles needed to understand these systems. In this section, we briefly overview two dynamical systems used in formulating generative models of this sort: Lotka-Volterra dynamics and Lorenz systems. The former may be used in characterizations of systems with a sequential aspect to their dynamics, while the latter represent chaotic systems.

Lotka-Volterra dynamics inherit from characterizations of predator-prey dynamics in ecology. While they have since found application in numerous disciplines, predator-prey systems remain a useful example to provide some intuition about their workings. When the predator population is small, the prey may increase their numbers to become a relatively large population. This provides additional food for the predators, whose population size then grows. Increased predation causes a decrease in the number of prey species and therefore a decrease in the number of predators. From here, the cycle continues. This gives an oscillatory pattern whereby the prey population size peaks, then the predators', then the preys' again, and so on. By generalizing this to more than two populations (e.g., carnivore, herbivore, and plant populations), we can generate a sequence of peaks. Figure 8.2 illustrates generalized Lotka-Volterra dynamics with three populations, which obey dynamics of the following form:

$$f(x, v) = x \circ (v + \mathbf{A}x) \quad (8.6)$$

Here, x is a vector as before. The \circ symbol means an elementwise product. Intrinsic birth and death rates are given by the vector v , and \mathbf{A} is a matrix whose elements are positive if the species indexed by the column prey on those indexed by the row and negative if the relationship is inverted.

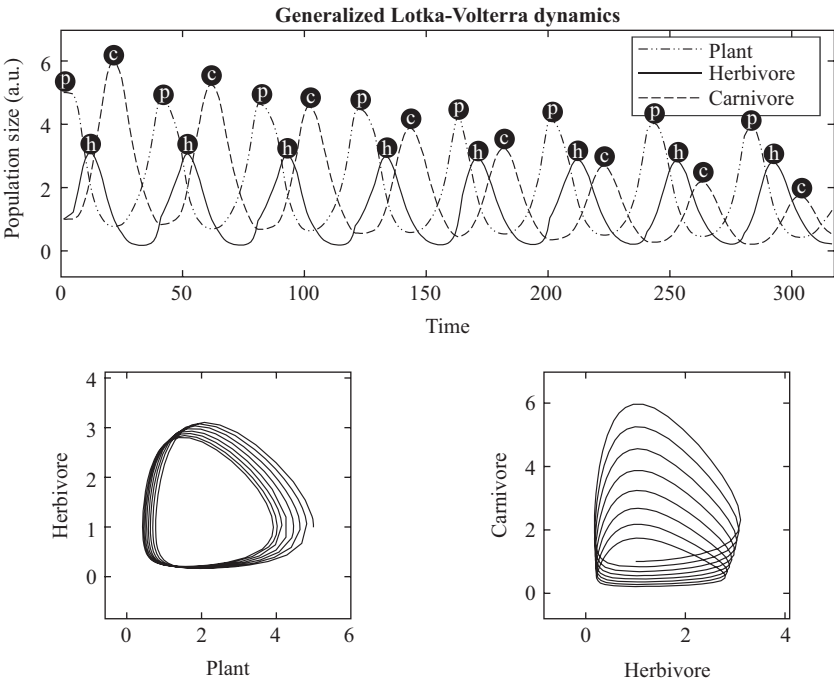


Figure 8.2

Generalized sequential dynamics emerging from Lotka-Volterra systems provide an important point of connection with the discrete sequential dynamics assumed in chapter 7. These dynamics can be applied to a range of systems but are framed here in terms of predator-prey relationships for ease of interpretation. *Top*: Population changes over time. The population size is expressed in terms of arbitrary units (a.u.). The peaks are labeled on the basis of which species has the greatest population at that point. The repeated pattern of *p*, *h*, *c* can be seen as a sequence of three (not necessarily evenly spaced) discrete time steps. *Bottom*: Trajectories emphasizing the (approximately) periodic pattern that each follows.

Figure 8.2 makes clear that having a generative model that incorporates Lotka-Volterra dynamics allows for temporal sequencing (Huerta and Rabino- vich 2004)—depending on the current highest peak. Each line can be thought of as representing a hidden state, in place of a species. Figure 8.3 highlights two important examples in which these dynamics have been exploited to generate behavior. One is a hierarchical model that uses the sequential dynamics afforded by a Lotka-Volterra system to time the response of an eye blink relative to a conditioned stimulus (Friston and Herreros 2016). The

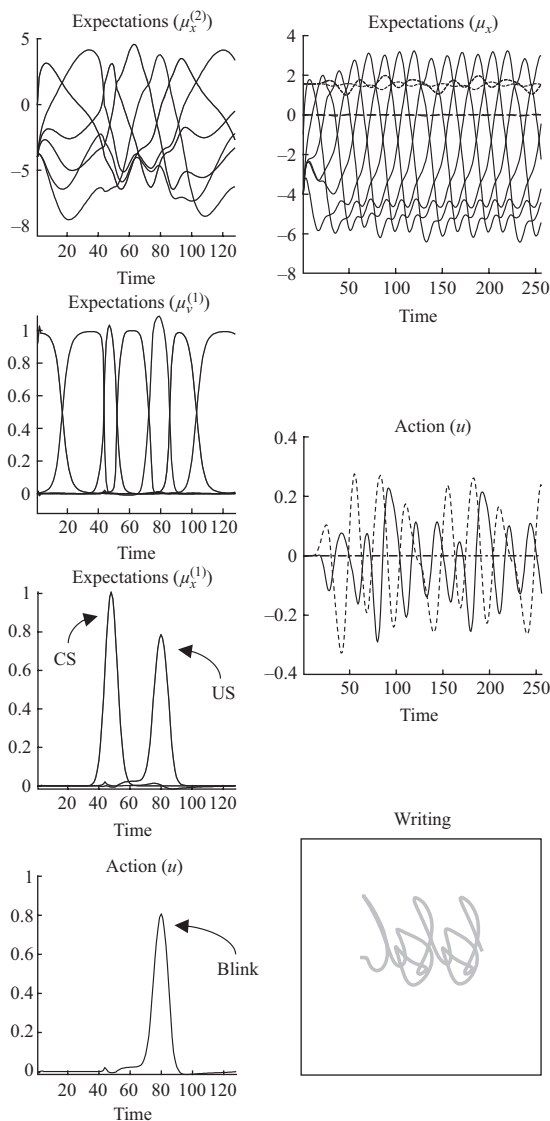


Figure 8.3

Two applications of generalized sequential Lotka-Volterra dynamics in Active Inference. *Left:* Eyeblink conditioning used to empirically investigate cerebellar function (Friston and Herreros 2016). Starting at the highest level of the column, the expected states show the same kind of sequential pattern as in figure 8.2. This passes down to the next level to predict sequential hidden causes; the various peaks here predict states at the next level down, where the first peak is the conditioned stimulus (CS) and the second is the unconditioned stimulus (US). Finally, the predicted US induces action—a blink. *Right:* Sequential peaks using an attracting point as in equation 8.5 but selecting the specific attractor on the basis of which population of a Lotka-Volterra system is currently highest; this leads to a sequential visiting of each point, giving rise to a form of handwriting (Friston, Mattout, and Kilner 2011).

paradigm is based on those used in the investigation of cerebellar function. An unconditioned stimulus (a puff of air directed toward an animal's eye) elicits a response (blinking). A conditioned stimulus (an auditory tone) may be played prior to the unconditioned stimulus on multiple occasions. By learning (see box 8.2) the number of peaks in the Lotka-Volterra dynamics that separate the conditioned stimulus from the unconditioned stimulus, the animal learns to preempt the air puff and time the appropriate blink. This is a form of temporal learning, since the number of peaks provides an implicit estimate of the length of the temporal interval from the conditioned to the unconditioned stimuli. In the second example in figure 8.3, each sequential peak is associated with an alternative attracting point that drives movements to a series of attracting points arranged to suggest hand-writing (Friston, Mattout, and Kilner 2011). As the two examples illustrate,

Box 8.2

Learning in continuous models

As discussed in chapter 7, learning is the process of optimizing beliefs about the parameters (θ) of a generative model. In the continuous-time domain, this means accumulating evidence over time. This works as if we treat data in a series of infinitesimally small time-intervals as obeying i.i.d. (independent and identically distributed) assumptions and formulate a generative model that generates observations from (time-invariant) parameters:

$$\begin{aligned}\ln p(\tilde{y}, \theta) &= \ln p(\theta) + \int \ln p(y(t) | \theta) dt \\ &\approx \ln p(\theta) - \int F[y(t) | \theta] dt\end{aligned}$$

This may be used to formulate a functional (S) that plays the role of a free energy for parameters using the time integral of the free energy conditioned on parameters. Using a Laplace approximation, we get the following, wherein α acts to accumulate free energy gradients (i.e., evidence gradients):

$$\begin{aligned}S(\theta) &= E_{q(\theta)}[\ln q(\theta) + \int F[y(t) | \theta] dt - \ln p(\theta)] \\ &\approx \int F[y(t) | \mu_\theta] dt - \ln p(\mu_\theta) \\ \dot{\mu}_\theta &= \partial_{\mu_\theta} S(\mu_\theta) \\ &= \partial_{\mu_\theta} \ln p(\mu_\theta) - \int \partial_{\mu_\theta} F[y(t) | \mu_\theta] dt \\ &= \partial_{\mu_\theta} \ln p(\mu_\theta) - \alpha \\ \dot{\alpha} &= \partial_{\mu_\theta} F[y(t) | \mu_\theta]\end{aligned}$$

generalized Lotka-Volterra systems afford useful models of sequential dynamics using a continuous dynamical system.

The POMDP formulation of chapter 7 has largely superseded the use of generalized Lotka-Volterra systems in Active Inference applications. However, it is useful to bear this kind of dynamic in mind as a plausible continuous system that might underwrite the discrete sequential dynamics of chapter 7. In addition, Lotka-Volterra systems make explicit the distinction between representations of sequences involved in temporally deep planning and representations of rates of change in generalized coordinates of motion (see chapter 4). Each has its place but deals with different sorts of problems.

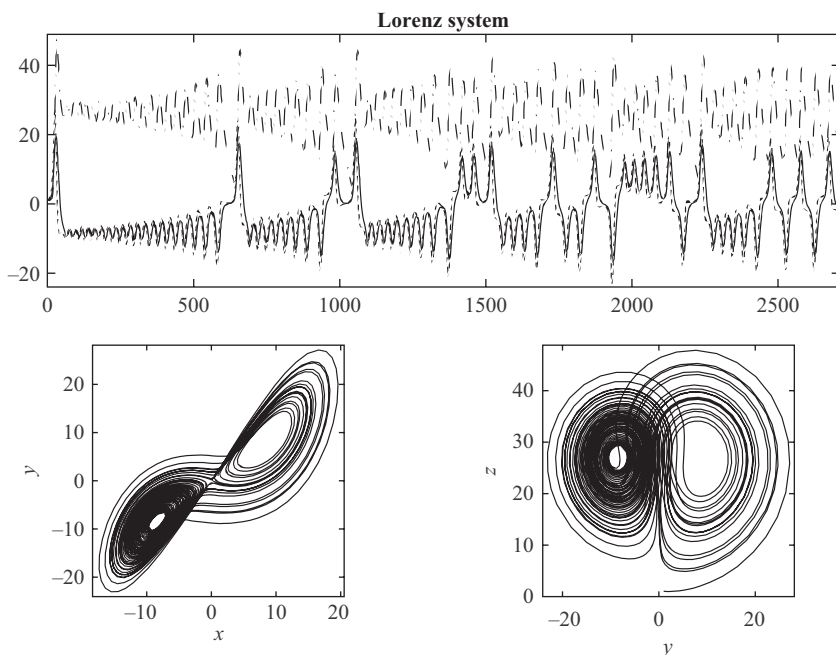
The second sort of dynamical system that has found widespread application in active inference research is the Lorenz system:

$$\dot{\mathbf{x}} = \begin{bmatrix} \sigma(x_2 - x_1) \\ x_1(\rho - x_3) - x_2 \\ x_1x_2 - \beta x_3 \end{bmatrix} \quad (8.7)$$

The parameters are known as the Prandtl number (σ), the Rayleigh number (ρ), and a constant (β) that relates to the physics of the system. Depending on the values these take, the system may behave in very different ways. Lorenz attractors were initially formulated to account for atmospheric convection dynamics. Their itinerant (wandering) behavior has prompted their use in generative models to simulate challenging inference problems. An important example of this is in the simulation of birdsong, which we unpack in the next section. These systems have also been used to simulate simple physical systems and to investigate the conditions under which their behavior starts to appear sentient. Figure 8.4 shows how the Lorenz system behaves under example parameter settings.

8.4 Generalized Synchrony

As mentioned above, a key example application of continuous state-space models is in a series of studies based on synthetic birdsong (Friston and Frith 2015b). An important aspect of these studies looks at communication and multi-agent inference problems. The idea here rests on the capacity of a creature to synchronize its internal states with something out there in the world (i.e., inference). When what is out there is another creature with a similar model, this synchronization means the internal states of one

**Figure 8.4**

Behavior of a Lorenz system attractor (using the same format as figure 8.2), showing how this 3-dimensional system evolves. Characteristically, it appears chaotic and unpredictable, spending some of its time orbiting one part of space before switching to another orbit. This itinerancy and apparent autonomy make this interesting system well suited to inclusion in models of biological phenomena.

creature should come to resemble the internal states of the other: a primitive kind of *theory of mind*.

Figure 8.5 shows the kind of generative model used to simulate songbirds. In this hierarchical model, high-level states (level 2) evolve according to a slow Lorenz system. One dimension of this system is then used to parameterize the Rayleigh number of a faster Lorenz system at the lower level (level 1). The lower-level variables then map to sensory (sonographic) data. Analogous to figure 8.1, the generative process additionally includes action; here, instead of moving a limb, actions influence the larynx, such that the sonographic data may be influenced by the bird. As before, action is generated to resolve prediction error. This means that if a bird hears the song it is predicting, there is no need to generate it itself. However, if it predicts a song that is not heard, it must start singing to resolve any error.

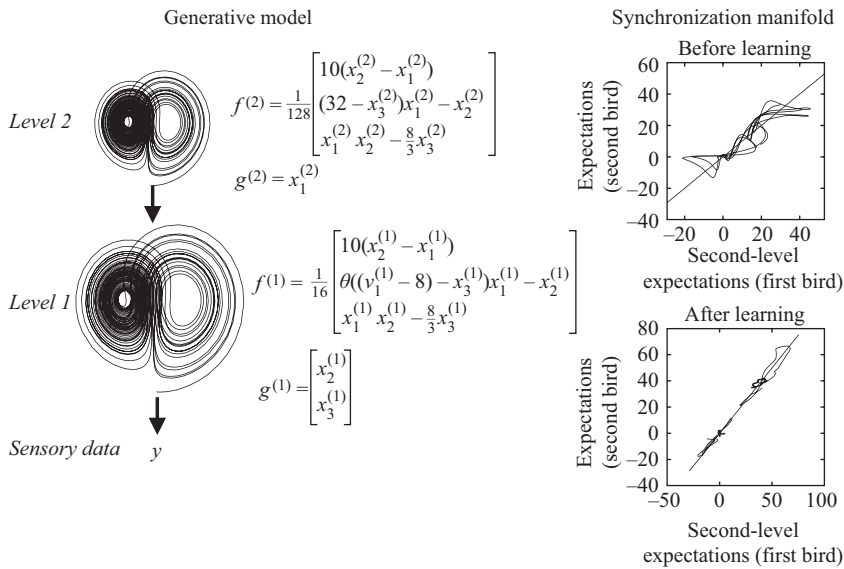


Figure 8.5
Synchronization and communication. *Left:* Generative model underwriting the bird-song simulations described in the main text. This is a hierarchical model, with Lorenz attractors at each level. *Right:* Synchronization manifolds of expectations at the second level for two birds before and after they have learned about one another. After learning the parameters of each other's generative models, the two bird's joint trajectory is confined to an (almost) 1-dimensional subspace, indicating synchronization.

This dynamic becomes more interesting when there are two birds in play, with similarly structured generative models. As long as one bird is singing, the other does not need to, as there is no error to resolve. However, if one bird stops singing, the other needs to continue the same song. This leads to a form of turn taking, sometimes phrased as “singing from the same hymn sheet,” with each bird contributing sections of the same song. What leads to this turn taking? Why doesn't one bird continue singing the whole song to its conspecific? The answer relates to the issue of sensory attenuation (see box 8.1), as acting to generate birdsong requires a reduction in the precision of predictions about the consequences of action. Just as in saccadic eye movements, this implies alternation between attention to sensory (visual or auditory) data and attenuation during (saccadic or vocal) action designed to change those data. When there are two agents involved, this leads to an

alternation between listening to the other and singing—a simple form of conversation.

For this synthetic conversation to work, it is essential that the two birds synchronize with one another and know where they are in the song (or conversational trajectory). This implies that the inferences about hidden states in the generative model should be aligned between the birds. On the upper right of figure 8.5, we show a synchronization manifold of two birds who have not yet optimized their generative models in relation to one another; this plots a trajectory of the beliefs each bird has about the higher-level hidden states. Synchronization implies that when one bird infers a specific hidden state value, the other bird should infer the same; therefore, we would expect the trajectory to stay fixed to the $x = y$ line (technically called *identical synchronization of chaos*). Fluctuations around this line imply imperfect synchronization, as this plot shows. After exposure to one another and learning the parameters of each other's generative models (see box 8.2), the synchronization is nearly perfect (lower-right plot). The implication is that each bird has learned about the other and is able to infer what is going on in the other's head. In short, they have learned to share the same narrative and “sing from the same hymn sheet.”

A more general form of synchronization does not require synchronization along the $x = y$ line. In generalized synchronization, the joint behavior occupies a lower, 1-dimensional space than the higher, 2-dimensional space that could be occupied by this behavior. However, this low-dimensional space (the synchronization manifold) may be curved or have some other shape; this is analogous to the 2-dimensional space we occupy on the surface of the planet, despite the surface being curved into a 3-dimensional sphere. In addition to its central role in social behavior, generalized synchronization—occupancy of a low-dimensional region of a high-dimensional joint space—is very important in characterizations of biological systems as engaging in inference (generalized synchrony between internal and external states). While we do not have the space to unpack this extensive subject here, the inferential perspective speaks to the failure of generalized synchrony associated with neuropsychiatric syndromes like autism. This kind of synchrony is important not only in continuous-time models but also in POMDP models of linguistic communication between multiple agents (Friston, Parr et al. 2020).

8.5 Hybrid (Discrete and Continuous) Models

As we have seen in this and the previous chapter, discrete and continuous models both have important applications in Active Inference. While many settings call for one or the other, a more holistic perspective acknowledges that both are likely in play. This means we need a way to combine these generative models so that a single model includes both continuous and discrete variables (Friston, Parr, and de Vries 2017). Such hybrid or mixed models allow inferences about sequential action plans and translations of these decisions into movements through a continuous model. Figure 8.6 shows the form of these models, with a POMDP at the higher level, behaving as described in chapter 7, which generates a continuous model of the

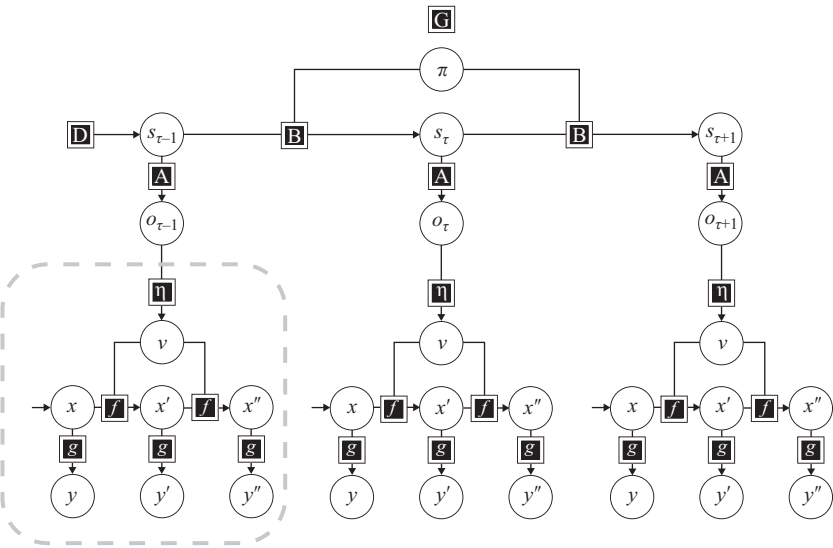


Figure 8.6 Mixed generative models in the form of a hierarchical model much like that in figure 7.12. However, there is an important difference between the form of the model at the higher level and that at the lower level. While the lower-level model (one example is highlighted by the dashed box) is the same form as the other models considered in this chapter—i.e., it is framed in terms of continuous states and continuous time and uses generalized coordinates of motion—the higher-level model is a POMDP model of the sort we saw throughout chapter 7—i.e., it is framed in terms of discrete states and times. Effectively, this means we can select (at regular time intervals) between alternative segments of a continuous trajectory.

sort addressed in this chapter at each discrete time step. This decomposes continuous time into a discrete sequence of short continuous trajectories.

To translate from the outcomes of the discrete level to the continuous level, we need to associate each alternative outcome with a point in some continuous space. To develop intuition for this idea, we consider the example of a delay-period oculomotor task—often used in primate research; see figure 8.7. This task involves three stages. First, a target appears in one of (for example) four possible locations, while a monkey maintains fixation on a central fixation cross. Next, the target disappears and must be remembered during a delay period. Finally, a signal is given for the monkey to make a saccade, at which point it must look at the location where the original target appeared. To complete this task, the monkey must be able to draw inferences about sequences (which stage of the task is currently in play) and to infer which of four locations to aim for. These are categorical inference problems suited to a POMDP formulation. However, once the appropriate location has been selected, the monkey must perform the eye movement that brings its fovea to the (continuous) coordinates of the target location.

Figure 8.7 illustrates the functioning of a mixed generative model that solves this problem. In the top panel, the model's higher level makes categorical decisions: it computes the posterior beliefs about four discrete target locations at four time periods. In the middle and bottom plots, the model's lower level computes continuous behavioral trajectories (eye movements) resulting from discrete inferences at the higher level.

Transforming decisions about discrete target locations into continuous eye movements requires each discrete target location (o) to be associated with a distribution over continuous hidden causes (v), which identifies the target coordinates. The prior over target coordinates can then be computed by taking the Bayesian model average over these locations, weighted by the inferences at the POMDP level:

$$\begin{aligned}
 P(\tilde{v} | o_\tau) &= \mathcal{N}(\tilde{\eta}_{o_\tau}, \tilde{\Pi}_v) \\
 P(\tilde{v}) &\approx \mathcal{N}(\tilde{\eta}, \tilde{\Pi}_v) \\
 \tilde{\eta} &= \mathbb{E}_{Q(o_\tau)}[\tilde{\eta}] = \mathbf{o}_\tau \cdot \tilde{\eta}
 \end{aligned} \tag{8.8}$$

To infer which discrete target best explains continuous data, we need to be able to compute the evidence associated with each hypothetical target—which is a function of observed continuous data. This exemplifies the fact that mixed models require reciprocal interactions between higher and lower hierarchical levels. As we see in figure 8.7, this facilitates the formation of

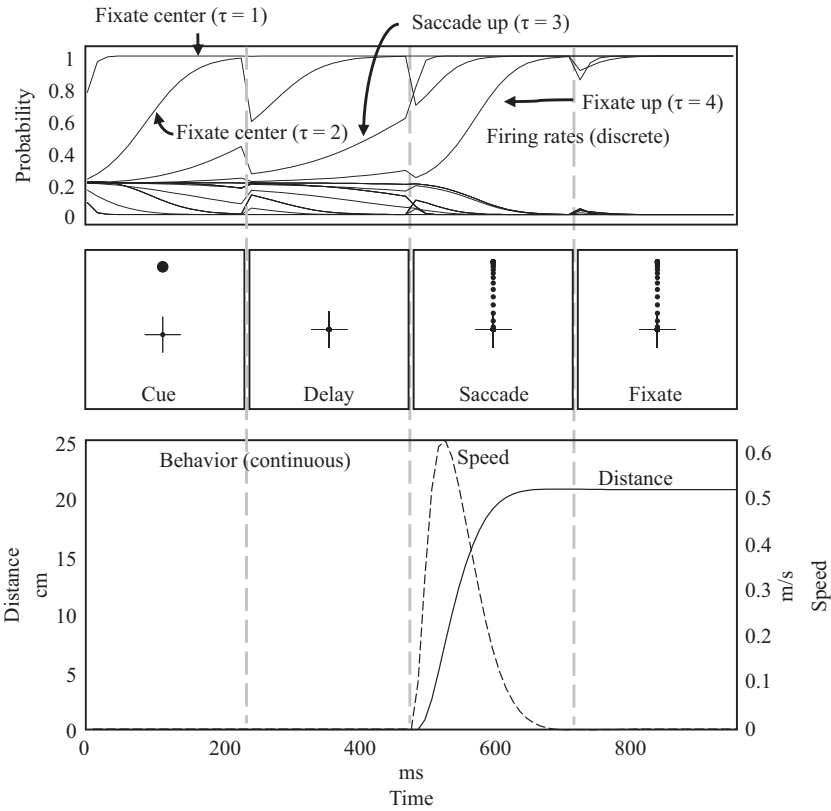


Figure 8.7

Transforming decisions into movement using mixed or hybrid models (Parr and Friston 2019b). This simple example uses the oculomotor delay-period paradigm outlined in the main text. *Top:* Neuronal firing rates representing posterior beliefs about the target location. The target may be in four different locations, and there are four time steps in this synthetic experiment, so there are 16 neural populations representing each of these combinations. The lines corresponding to the final inferred state of affairs are annotated explicitly. Note the belief updating at the first time step (from 0 to 250 ms), when the target initially appears, and at the third time step (from 500 to 750 ms), when the agent observes itself performing a saccade to that location. *Middle:* Behavior (i.e., a saccade from the center location to the target location). During the first quarter of the experiment, the target is visible in the upper location. Next, there is a delay period, during which fixation is maintained. Then a saccade is performed to the correct location. Finally, there is a period during which fixation is maintained at the target location. *Bottom:* Continuous behavioral trajectories that result from the discrete inferences of the top plot.

Box 8.3

Mixture models and clustering

The issue of combining categorical and continuous generative models outside of Active Inference has primarily been framed through the lens of clustering. Here, the aim is to assign each (continuous) data point to a (discrete) cluster. A range of algorithms have been employed to solve this problem, but most of them implicitly rely on a generative model similar to that used here. This is a *mixture of Gaussians* (aka a *Gaussian mixture model*):

$$P(\tilde{y}, \tilde{s}, \mathbf{D}, \eta, \Pi) = P(\mathbf{D})P(\eta)P(\Pi) \prod_i P(s_i | \mathbf{D})P(y_i | s_i, \eta, \Pi)$$

$$P(\mathbf{D}) = \text{Dir}(d)$$

$$P(s_i | \mathbf{D}) = \text{Cat}(\mathbf{D})$$

$$P(y_i | s_i, \eta, \Pi) = \mathcal{N}(\eta_{s_i}, \Pi_{s_i})$$

The problem in clustering approaches is to infer the mean and precision (η and Π , respectively) of each cluster and the posterior probability that each data-point (y_i) belongs to a given cluster $P(s_i | y_i)$. For our purposes (as described in section 8.5), we assume precise (delta-function) priors for η and Π and calculate $Q(s_i) \approx P(s_i | y_i)$ via Bayesian model reduction (see box 7.3).

beliefs about where to perform saccades and the execution of those saccades at the appropriate times. At the first discrete time step (up to 250 ms on the continuous scale), the monkey is able to infer with some confidence that its eyes are centered on the fixation cross during the first time step, that it will maintain this fixation at the second time step (up to 500 ms), and that the most likely course of action after this will result in foveating the upper location. This can be seen in the discrete firing rates (top plot). This translates into the continuous behavior which, when implemented, increases the confidence in beliefs about the discrete states (note the increase in probability for an upward saccade at the third time step once the continuous data become available between 500 and 750 ms). This simple example, based on experimental cognitive research, illustrates the basic principles of translation between discrete action plans and their continuous implementation.

8.6 Summary

In this chapter, we have overviewed the applications of continuous-time generative models under Active Inference. This is a huge topic, and much has been left out (see table 8.1 for further reading). However, the broad

Table 8.1

Key advances in continuous-time models

Application	Sources	Notes
Synthetic birdsong	Friston and Frith 2015a Friston and Frith 2015b Isomura, Parr, and Friston 2019	This series of papers deals with communication and the interaction between synthetic agents, a simulated pair (or group) of songbirds singing to one another. The studies unpack phenomena from generalized synchrony to perceptual inference to sensory attenuation.
Oculomotor delays	Perrinet, Adams, and Friston 2014	By taking advantage of beliefs about the near past and future implicit in models formulated in generalized coordinates of motion, it is possible to account for sensorimotor delays through projections a short way into the future or past.
Conditioned reflexes	Friston and Herreros 2016	Using a model based on a Lotka-Volterra system, the temporal relationship between a conditioned and unconditioned stimulus is learned and used to generate an anticipatory blink.
Smooth pursuit eye movements	Adams, Perrinet, and Friston 2012	This work looks at the role of smooth pursuit eye movements, following a visual target. It aims to reproduce differences between neurotypical and schizophrenic individuals in response to pursuit with and without visual occlusion.
Psychosis	Adams, Stephan et al. 2013	Building on the songbird and smooth pursuit models, this research looks at how false, psychotic inference may arise from suboptimal prior beliefs.
Illusions	Brown and Friston 2012 Brown, Adams et al. 2013	Illusions offer a useful tool to reveal the prior beliefs our brains appeal to in the presence of uncertain or ambiguous sensory input. These papers take several examples of common illusions and demonstrate the optimality of illusory inferences under certain prior beliefs.

Table 8.1
(continued)

Application	Sources	Notes
Saccades	Friston, Adams et al. 2012 Donnarumma et al. 2017 Parr and Friston 2018a	Like the smooth pursuit simulations, these papers consider eye movement control. However, here the eyes do not simply follow a target but must move to one of several possible target locations. They deal with the generative models we need to be able to do this and (once the models have been specified) the emergent architectures and physiology.
Action observation	Friston, Mattout, and Kilner 2011	This work considers the role of the mirror neuron system and formalizes the idea that generative models of our own actions can also be put to use in modeling, and replicating, behavior observed in others.
Attention	Feldman and Friston 2010 Kanai et al. 2015	Through predicting precision, we implicitly select the data that we believe is most informative. This work highlights how implementations of this idea reproduce classical psychophysical findings in the Posner paradigm and figure-ground discrimination tasks.
Hybrid models	Friston, Parr, and de Vries 2017 Parr and Friston 2018c Parr and Friston 2019b	These models make use of discrete POMDP models in combination with predictive coding schemes. Most current examples of this modeling are framed in terms of visual search behavior or oculomotor control. These require selecting where to look and then implementing the process of looking there.
Self-organization	Friston 2013 Friston, Levin et al. 2015 Palacios et al. 2020	This line of research is based on the idea that groups of cells can organize into a predefined structure when each cell has the same implicit generative model of that structure. Specifically, they must know what sort of sensory input they would predict if they were a particular kind of cell.

concepts outlined here provide a foundation from which these models may be further explored. Specifically, we have considered movement generation in terms of the fulfillment of predictions. This greatly simplifies the treatment of motor control problems, as we do not need any additional machinery or inverse models—just spinal or brain stem reflex arcs. We highlighted the role of precision and sensory attenuation in motor control of this sort. Given that a key advantage of continuous schemes is to articulate generative models in terms of dynamical systems, we outlined two ubiquitous types of dynamical system that have found widespread application in Active Inference research. Generalized Lotka-Volterra systems act to provide temporal sequencing in a continuous context, while Lorenz attractors may be used to generate rich simulations, including synthetic birdsong. Next we considered the concept of generalized synchrony. Synchronization of the internal states of a system to external states forms the basis of inferential treatments of brain function and is crucial in accounts of social systems—where external states largely comprise conspecifics (i.e., creatures like me). Finally, we set out the unification of the discrete and continuous models of chapters 7 and 8, bringing together the expected free energy minimizing (exploitative and explorative) dynamics of POMDP formulations, the enaction of the behaviors these mandate through continuous processes, and the reciprocal message passing that mediates this interaction. In short, this takes us from decisions to movements—and back again.

This is a section of [doi:10.7551/mitpress/12441.001.0001](https://doi.org/10.7551/mitpress/12441.001.0001)

Active Inference

The Free Energy Principle in Mind, Brain, and Behavior

By: Thomas Parr, Giovanni Pezzulo, Karl J. Friston

Citation:

Active Inference: The Free Energy Principle in Mind, Brain, and Behavior

By: Thomas Parr, Giovanni Pezzulo, Karl J. Friston

DOI: 10.7551/mitpress/12441.001.0001

ISBN (electronic): 9780262369978

Publisher: The MIT Press

Published: 2022

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

© 2022 Massachusetts Institute of Technology

This work is subject to a Creative Commons CC BY-NC-ND license.
Subject to such license, all rights are reserved.



The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data is available.

Names: Parr, Thomas, 1993– author. | Pezzulo, Giovanni, author. | Friston, K. J. (Karl J.), author.

Title: Active inference : the free energy principle in mind, brain, and behavior / Thomas Parr, Giovanni Pezzulo, and Karl J. Friston.

Description: Cambridge, Massachusetts : The MIT Press, [2022] | Includes bibliographical references and index.

Identifiers: LCCN 2021023032 | ISBN 9780262045353 (hardcover)

Subjects: LCSH: Perception. | Inference. | Neurobiology. | Human behavior models. | Knowledge, Theory of. | Bayesian statistical decision theory.

Classification: LCC BF311 .P31366 2022 | DDC 153—dc23

LC record available at <https://lcn.loc.gov/2021023032>