Notes

Chapter 1

1. The term *normative* means there is some evaluative standard against which behavior can be scored. Active Inference is normative in the sense that perception and action are scored by free energy—a quantity we will unpack throughout this and the next chapters.

2. *Bayes optimality* refers to a set of related concepts that deal with aspects of Bayes' theorem—something we will unpack in chapter 2. Broadly, it refers to any action that minimizes (or maximizes) the expected value of some cost (or utility) function given some observation. This encompasses Bayes-optimal experimental design, wherein an experiment (action) is chosen to maximize expected information gain.

Chapter 2

1. Like bits, nats are units of information. The choice of unit depends on whether we use a logarithm to the base 2 (bits) or a natural logarithm (nats).

2. *Support* is a technical term referring to the possible arguments for a distribution. For example, the support of a categorical probability distribution is a series of alternative states (i.e., event space) whose probability may be quantified. The support of a univariate normal distribution is the entire real number line.

3. The details of this table are not important for understanding Active Inference conceptually, but for interested readers, we briefly unpack the key points. The Support column tells us the set of variables whose surprise can be quantified using each distribution. This is the set of real numbers for the Gaussian distribution. For the multinomial distribution, the support comprises a group of K variables, each taking an integer value up to a maximum N, under the constraint that all elements in that group sum to N. For the Dirichlet distribution, the support includes any group of K real numbers between 0 and 1, where all elements in the group sum to 1. The gamma distribution quantifies the surprise of nonnegative real numbers. The Surprise column shows the

way in which the surprise can be calculated. This depends on constants (in addition to the random variable x) that control the shape of the underlying distribution.

4. Interestingly, resource limitations are not the only barrier to exact Bayesian inference. In the presence of complex models, exact inference may be analytically intractable, such that no additional resources could help solve the exact problem.

5. Like the KL-Divergence, entropy is a quantity from information theory. It is a measure of the dispersion (or uncertainty) of a probability distribution. Technically, it is the average of the negative log probability or average surprise.

6. *Complexity* as used here scores the degree to which we must depart from our prior beliefs about the world in order to explain data.

7. This is referred to as *accuracy* because an explanation's accuracy increases when a high log probability of outcomes, expected under the inferred hidden states, is assigned to observed data—i.e., when the predicted distribution of outcomes accurately captures the measured distribution.

Chapter 3

1. *Nonequilibrium* here refers to the absence of *detailed balance*. Detailed balance is the invariance of a system under time reversal once it has reached steady state. We can see that the system on the left of figure 3.3 does not possess detailed balance, as the trajectory tends to curve counterclockwise around the contours of surprise. If we were to play this back in reverse, the system would appear to rotate clockwise.

2. This is not the same as saying that surprise-minimizing systems must minimize their entropy. As we see in figure 3.3, the system does not tend toward an infinitely precise (point) distribution that would minimize entropy, but it maintains a consistent dispersion over time—bounding entropy from above and below.

3. The capital A is used to distinguish Action as a path integral of a Lagrangian from action as the dynamics of active states of a Markov blanket.

4. A Lagrangian is a function of a position and velocity that gives the difference between kinetic and potential energies. A Hamiltonian is related to (via a Legendre transform) and expresses the total energy of the system in terms of position and momentum.

Chapter 4

1. Here and throughout the chapter, the conditioning on the model is left implicit; hence, the model evidence is written as $P(y)$ and not $P(y|m)$.

2. Technically, this is true for any concave function, but we are concerned only with logarithms here.

3. An expectation is a weighted sum or integral of the term inside the square brackets; each term is weighted by the probability indicated by the subscript (see box 2.2).

4. In this book, we follow the physicist's convention in which the free energy is an upper bound on the negative log evidence. However, other disciplines (including statistics and machine learning) use the negative free energy as an evidence lower bound (or ELBO). These are completely equivalent but can cause some confusion in interdisciplinary research.

5. MAP estimates are the most probable states considering prior beliefs and the data available; contrast this with *maximum likelihood* approaches which do not take prior beliefs into account.

Chapter 5

1. This nomenclature comes from reinforcement learning theories (Daw et al. 2005) but is slightly misleading as both systems depend on models. "Model-free" systems just use a simpler model that predicts a certain kind of behavior in a certain kind of environment.

Chapter 6

1. This does not imply discrete temporal dynamics from a neural perspective. Instead, continuous neural dynamics are seen as representing (continuous) changes in beliefs about (discrete) sequences of events.

2. Having said this, the use of generalized coordinates of motion (box 4.2) in continuous-time models means that they are temporally deep in virtue of their implicit representation of a short trajectory. However, these models do not (necessarily) include variables representing alternative trajectories one could pursue (i.e., the consequences of sequences of actions).

3. Do not confuse temporally deep models with hierarchical models. Unlike temporally deep models, some hierarchical models (e.g., predictive coding models; see section 4.4.1) only consider present observations. However, generative models can be both hierarchical and temporally deep to afford multiscale planning.

Chapter 8

1. Often it is necessary to add damping terms to account for friction and/or viscosity to preclude oscillatory solutions.

Chapter 9

1. Practically, it is often useful to define parameters as log scaling parameters: the parameter acts as a nonnegative scaling factor and cannot be characterized by a normal distribution, which allocates negative numbers a finite probability density. Assuming instead that the log of the scaling parameter is normally distributed ensures positivity when exponentiated to get the scaling parameter itself. The same aim may be achieved by modeling the square root of a parameter as being normally distributed.

2. For example, $\partial_x f(x) \approx \frac{1}{2\Delta x}\big(f(x+\Delta x) - f(x-\Delta x)\big)$.

Chapter 10

1. From a more pragmatic viewpoint, Active Inference only requires the acquisition of forward models, which are (typically) easier to learn compared to inverse models because they are simply a direct (observable) mapping between actions and consequences. Forward models can also be acquired by imitation or external supervision—a technique largely analogous to Active Inference that is widely used to train robotic models (Nishimoto and Tani 2009).

2. In machine learning, the process of optimizing sequences of actions is sometimes called *sequential policy optimization*—as opposed to the more usual *optimization of state-action policies*—namely, "If I am in this state, what do I do?"

3. The notion of deploying cognitive resources efficiently is an inherent part of free energy minimization because minimizing complexity automatically maximizes efficiency, in both an information theoretic and thermodynamic sense. Put simply, the path of least resistance is the path of least free energy.

Appendix A

1. Tensors are a generalization of the concepts of scalars, vectors, and matrices. Heuristically, we can think of these as arrays whose elements are addressable by a certain number of indices. For a vector, we need only a single (row) argument to specify an element. This makes it a first order tensor. For a matrix, we need to specify a column and a row, making it a second order tensor. Scalars need no indices to specify an element, so are 0 order.

2. This uses the identity $\partial_A \ln |A| = A^{-1}$.

3. In the context of variational inference, the integral is typically an expectation.

4. This is sometimes referred to as *the fundamental lemma of variational calculus*.

5. This uses the chain rule, as applied to the derivative of a log: $\partial_x \ln f(x) = f(x)^{-1} \partial_x f(x)$.

Appendix B

1. A normalized exponential function.

2. For concision, we have omitted some terms in the derivatives of the log partition functions. We are licensed to do so by the choice of variational distribution, as any higher order polynomial terms would violate the form of this distribution.

3. The C here is a covariance and should not be confused with a prior preference, despite the same notation in preceding sections.