6   A Recipe for Designing Active Inference Models

Give me six hours to chop down a tree and I will spend the first four sharpening the axe.
—Abraham Lincoln

6.1   Introduction

This chapter provides a four-step recipe to construct an Active Inference model, discussing the most important design choices one has to make to realize a model and providing some guidelines for those choices. It serves as an introduction to the second part of the book, which will illustrate several specific computational models using Active Inference and their applications in a variety of cognitive domains.

As Active Inference is a normative approach, it tries to explain as much as possible about behavior, cognitive, and neural processes from first principles. Consistently, the design philosophy of Active Inference is top-down. Unlike many other approaches to computational neuroscience, the challenge is not to emulate a brain, piece by piece, but to find the *generative model* that describes the problem the brain is trying to solve. Once the problem is appropriately formalized in terms of a generative model, the solution to the problem emerges under Active Inference—with accompanying predictions about brains and minds. In other words, the generative model provides a complete description of a system of interest. The resulting behavior, inference, and neural dynamics can all be derived from a model by minimizing free energy.

The *generative modeling* approach is used in several disciplines for the realization of cognitive models, statistical modeling, experimental data analysis,

and machine learning (Hinton 2007b; Lee and Wagenmakers 2014; Pezzulo, Rigoli, and Friston 2015; Allen et al. 2019; Foster 2019). Here, we are primarily interested in designing generative models that engender cognitive processes of interest. We have seen this design methodology in previous chapters. For example, using a generative model for predictive coding, perception was cast as an inference about the most likely cause of sensations; using a generative model that evolves in discrete time, planning was cast as an inference about the most likely course of action. Depending on the problem of interest (e.g., planning during spatial navigation or planning saccades during visual search), one can adapt the form of these generative models to equip them with different structures (e.g., shallow or hierarchical) and variables (e.g., beliefs about allocentric or egocentric spatial locations). Importantly, Active Inference may take on many different guises under different assumptions about the form of the generative model being optimized. For example, assumptions about models that evolve in discrete or continuous time influence the form of the message passing (see chapter 4). This implies that the choice of a generative model corresponds to specific predictions about both behavior and neurobiology.

This flexibility is useful as it allows us to use the same language to describe processes in multiple domains. However, it can also be confusing from a practical perspective, as there are a number of choices that must be made to find the appropriate level of description for the system of interest. In the second part of this book, we will try to resolve this confusion through a series of illustrative examples of Active Inference in silico. This chapter introduces a general recipe for the design of Active Inference models, highlighting some of the key design choices, distinctions, and dichotomies that will appear in the numerical analysis of computational models described in subsequent chapters.

6.2   Designing an Active Inference Model: A Recipe in Four Steps

Designing an Active Inference model requires four foundational steps, each resolving a specific design question:

1. **Which system are we modeling?** The first choice to make is always the system of interest. This may not be as simple as it seems; it rests on the identification of the boundaries (i.e., Markov blanket) of that system. What counts as an Active Inference agent (generative model), what

counts as the external environment (generative process), and what is the interface (sensory data and actions) between them?

2. **What is the most appropriate form for the generative model?** The first of the next three practical challenges is deciding whether it is appropriate to think of a process more in terms of categorical (discrete) inferences or continuous inferences, motivating the choice between discrete or continuous-time implementations (or a hybrid) of Active Inference. Then we need to select the most appropriate hierarchical depth, motivating the choice between shallow versus deep models. Finally, we need to consider whether it is necessary to endow generative models with temporal depth and the ability to predict action-contingent observations to support planning.

3. **How to set up the generative model?** What are the generative model's most appropriate variables and priors? Which parts are fixed and what must be learned? We emphasize the importance of choosing the right sort of variables and prior beliefs; furthermore, we emphasize a separation in timescales between the (faster) update of state variables that occurs during inference and the (slower) update of model parameters that occurs during learning.

4. **How to set up the generative process?** What are the elements of the generative process (and how do they differ from the generative model)?

These four steps (in most cases) suffice to design an Active Inference model. Once completed, the behavior of the system is determined by the standard schemes of Active Inference: the descent of the active and internal states on the free energy functional associated with the model. From a more practical perspective, once one has specified the generative model and generative process, one can use standard Active Inference software routines to obtain numerical results, as well as to perform data visualization, analysis, and fitting (e.g., model-based data analysis). In what follows, we will review the four design choices in order.

6.3   What System Are We Modeling?

A useful first step in applying the formalism of Active Inference is to identify the boundaries of the system of interest because we are interested in characterizing the interaction between what is *internal* to a system and the *external world* via sensory receptors and effectors (e.g., muscles or glands). As

discussed in chapter 3, a formal way to characterize the distinction between internal states of a system and external variables (and intermediate variables that mediate their interactions) is in terms of a Markov blanket (Pearl 1988). To reiterate the argument, a Markov blanket may be subdivided into two sorts of variables (Friston 2013): those that mediate the influence of the external world on internal states of the system of interest (i.e., sensory states) and those that mediate the influence of internal states of the system of interest on the external world (i.e., active states). See figure 6.1.

Importantly, there are many ways in which a boundary between internal and external may be defined. In most of the simulations we will discuss in the second part of this book, there will be a (Markov blanket) separation between an agent (roughly, a living organism) and its environment. This corresponds to the usual setup of cognitive models, where an agent implements cognitive processes such as perception and action selection on the basis of its internal (e.g., brain) states and is provided with sensors and effectors.
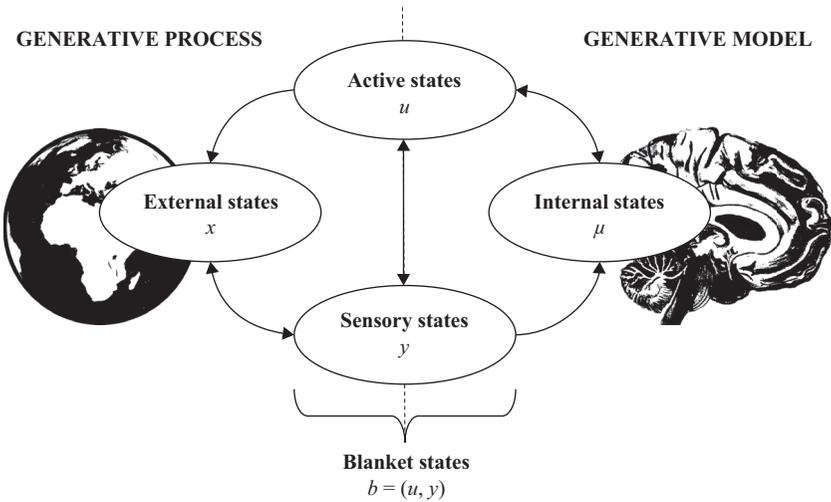
Figure 6.1
Action-perception loop between an adaptive system (here, the brain) and the environment, along with the Markov blanket (composed of active states and sensory states) that mediates their interaction. The figure implies that the adaptive system only affects the environment by performing actions (via active states) and that the environment only affects the adaptive system by producing observations (via sensory states). The figure exemplifies the distinction between the adaptive system's generative model and the (external) generative process that produces its observations.

However, this is not the only possibility. From the perspective of neurobiology, we could draw a Markov blanket around a single neuron, around the brain, or around the entire body. In the first case, sensory states include postsynaptic receptor occupancies, and active states include the rate at which vesicles containing neurotransmitters fuse with the presynaptic membrane. The internal states of the neuron (e.g., membrane potentials, calcium concentrations) can then be thought of as inferring the causes of its sensory states according to some (implicit) generative model (Palacios, Isomura et al. 2019). This setup treats the external states (that are being modeled) as including the neuronal network in which our neuron participates. This is very different from the inference taking place when we assume our entire network is internal to the Markov blanket. For example, if we take a system whose sensory states are the photoreceptors in the retina and whose active states are the oculomotor muscles, the inferences performed by the internal states are about things outside the brain. This speaks to the importance of scale, as the internal states of this Markov blanket include the internal states from the perspective of a single neuron. The latter internal states appear to make inferences about things within the brain when the Markov blanket is drawn around a single neuron but not when the blanket is drawn around the nervous system.

The above is particularly relevant when dealing with embodied or extended perspectives on cognition (Clark and Chalmers 1998; Barsalou 2008; Pezzulo, Lw et al. 2011). For example, if we draw the blanket around the nervous system, the rest of the body becomes an external state, about which we must make inferences from interoceptive sensory states (Allen et al. 2019). Alternatively, we could draw our blanket around the entire organism. This would make it look as if organs other than the brain were making inferences about their environment. For example, depression of the skin in response to an external pressure could be framed as an inference about the source of the external pressure. The extended cognition perspective takes this further and says that objects external to the body may be incorporated into the Markov blanket (e.g., the use of a calculator to assist in inference implies that the calculator is part of the internal state-space of the inferring system). Finally, we could have multiple Markov blankets, nested within one another (e.g., brains, organisms, communities).

In sum, defining the Markov blanket ensures we know what is being inferred (external states) and what is doing the inferring. Indeed, minimization of free

energy with respect to a generative model only involves the internal and active states of a system: these only see the sensory states, so they can only infer the external state of the world vicariously.

6.4 What Is the Most Appropriate Form for the Generative Model?

Once we have decided on the internal states of a system and the states that mediate their interaction with the world outside, we need to specify the *generative model* that explains how external states influence sensory states.

As discussed in previous chapters, Active Inference can operate on different kinds of generative models. Therefore, we need to specify the most appropriate form of the generative model for the problem at hand. This implies making three main design choices. The first is a choice between models that include continuous or discrete variables (or both). The second is a choice between shallow models, in which inference operates on a single timescale (i.e., all variables evolve at the same timescale), and hierarchical or deep models, in which inference operates on multiple timescales (i.e., different variables evolve at different timescales). The third is a choice between models that only consider present observations versus models having some temporal depth, which consider the consequences of actions or plans.

6.4.1 Discrete or Continuous Variables (or Both)?

The first design choice is to consider whether generative models that use discrete or continuous variables are more appropriate. The former include object identities, alternative action plans, and discretized representations of continuous variables. These are modeled through expressing the probability—at each time step—of one variable transitioning into another type. The latter include things like position, velocity, muscle length, and luminance and require a generative model expressed in terms of rates of change.

Computationally, the distinction between the two may not be clear-cut because a continuous variable may be discretized, and a discrete variable may be expressed through continuous variables. However, this distinction is important conceptually, as it underlies specific hypotheses about the time course (discrete or continuous) of the cognitive processes of interest.[1] In most current implementations of Active Inference, high-level decision processes, such as the choice between alternative courses of actions, are

modeled using discrete variables, whereas more fine-grained perception and action dynamics are implemented using continuous variables; we will provide examples of both in chapters 7 and 8, respectively.

Furthermore, the choice between discrete and continuous variables is relevant for neurobiology. While each style of modeling appeals to free energy minimization, the message passing these imply take different forms. To the extent that one considers message passing relevant for a process theory (see chapter 5), this implies that the neural dynamics that realize this minimization are different under each sort of model. Continuous schemes underwrite predictive coding—a theory of neural processing that relies on top-down predictions corrected by bottom-up prediction errors. However, the analogous process theories for discrete inferences involve messages of a different form. Finally, the two types of model may be combined such that discrete states are associated with continuous variables. This means we can specify a generative model wherein a discrete state (e.g., object identity) generates some pattern of continuous variables (e.g., luminance). We will discuss an example of a hybrid or mixed generative model that includes both discrete and continuous variables in chapter 8.

6.4.2   Timescales of Inference: Shallow versus Hierarchical Models

The second design choice concerns the timescales of Active Inference. One can select either (shallow) generative models, in which all the variables evolve at the same timescale, or (hierarchical or deep) models, which include variables that evolve at different timescales: slower for higher levels and faster for lower levels.

While many simple cognitive models only require shallow models, these are not sufficient when there is a clear separation of timescales between different aspects of a cognitive process of interest. One example of this is in language processing, in which short sequences of phonemes are contextualized by the word that is spoken and short sequences of words are contextualized by the current sentence. Crucially, the duration of the word transcends that of any one phoneme in the sequence and the duration of the sentence transcends that of any one word in the sequence. Hence, to model language processing, one can consider a hierarchical model in which sentences, words, and phonemes appear at different (higher to lower) hierarchical levels and evolve over (slower to faster) timescales that are approximately independent of one another. This is only an approximate separation, as levels must

influence each other (e.g., the sentence influences the next words in the sequence; the word influences the next phonemes in the sequence).

However, this does not mean we need to attempt to model the entire brain to develop meaningful simulations of a single level. For example, if we wanted to focus on word processing, we could address some aspects without having to deal with phoneme processing. This means we can treat input from parts of the brain drawing inferences about phonemes as providing observations from the perspective of word-processing areas. Phrasing this in terms of a Markov blanket, this typically means we treat the inferences performed by lower levels of a model as part of the sensory states of the blanket. This means we can summarize the inferences performed at the timescale of interest without having to specify the details of lower-level (faster) inferential processes—and this hierarchical factorization entails great computational benefits.

Another example is in the domain of intentional action selection, where the same goal (enter your apartment) can be active for an extended period of time and contextualizes a series of subgoals and actions (find keys, open door, enter) that are resolved at a much faster timescale. This separation of timescales, whether in the continuous or discrete domain, demands a hierarchical (deep) generative model. In neuroscience, one can assume that cortical hierarchies embed this sort of temporal separation of timescales, with slowly evolving states at higher levels and rapidly evolving states at lower levels, and that this recapitulates environmental dynamics, which also evolve at multiple timescales (e.g., during perceptual tasks like speech recognition or reading). In psychology, this sort of model is useful in reproducing hierarchical goal processing (Pezzulo, Rigoli, and Friston 2018) and working memory tasks (Parr and Friston 2017c) of the sort that rely on delay-period activity (Funahashi et al. 1989).

6.4.3 Temporal Depth of Inference and Planning

The third design choice concerns the temporal depth of inference. It is important to draw a distinction between two kinds of generative model: the first have *temporal depth* and represent explicitly the consequences of actions or action sequences (policies or plans), whereas the second lack temporal depth and only consider present but not future observations. These two kinds of model are exemplified in figure 4.3: the dynamic POMDP at the top and the continuous-time model at the bottom.[2] The key difference between these

two models is not that they use discrete or continuous variables, respectively, but that only the former (temporally deep) model endows creatures with the ability to plan ahead and select among possible futures.

Imagine a rodent who plans a route to a known food location in a maze. Doing this benefits from a temporally deep model, loosely equivalent to a spatial or cognitive map (Tolman 1948), which encodes contingencies between present and future locations conditioned on actions (e.g., the future location after turning right or left). The animal can use the temporally deep model to counterfactually consider multiple courses of action (e.g., series of right and left turns) and select the one expected to reach the food location.

Why is a temporally deep model required for planning? In Active Inference, planning is realized by calculating the expected free energy associated with different actions or policies and then selecting the policy that is associated with the lowest expected free energy. Expected free energy is not just a function of present observations (like variational free energy) but also a functional of future observations. The latter cannot be observed (by definition) but only predicted using a temporally deep model, which describes the ways in which actions produce future observations.

When designing an Active Inference agent it is useful to consider whether it should have planning and future-oriented capacities—and, in this case, to select a temporally deep model. Furthermore, it is useful to consider planning depth—that is, how far in the future the planning process can look. Finally, one can design generative models that are both hierarchical and temporally deep, wherein planning proceeds at multiple timescales—faster at lower levels, and slower at higher levels.[3] The decision whether to model alternative futures, contingent on policy selection, is largely tied up with the choice between discrete and continuous models because the idea of selecting between alternative futures, defined by sequences of actions, is more simply articulated using discrete-time models.

6.5   How to Set Up the Generative Model?

When we have specified our system of interest and identified the relevant forms of the generative model (e.g., continuous or discrete representation, shallow versus hierarchical structure), our next challenges are to specify the specific variables to include in the generative model and decide which of these variables remain fixed or change as an effect of learning.

6.5.1   Setting Up the Variables of the Generative Model

The variables of generative models can be either predefined or learned from data. For illustrative purposes, most models that we discuss in this book use predefined variables. When designing these models, in practice, the main challenge is deciding which hidden states, observations, and actions are most appropriate for the problem at hand. For example, the perceptual model able to distinguish frogs from apples in chapter 2 only included two hidden states (frogs, apples) and two observations (jumps, does not jump). A more sophisticated model could include additional observations (e.g., red, green) as well as actions such as touching, which produce differential sensory effects (jump or no jump) in the presence of a frog or an apple.

Figure 6.2 schematically illustrates a generative model for the concept of a jumping frog. The concept is cast as a hierarchical model, where
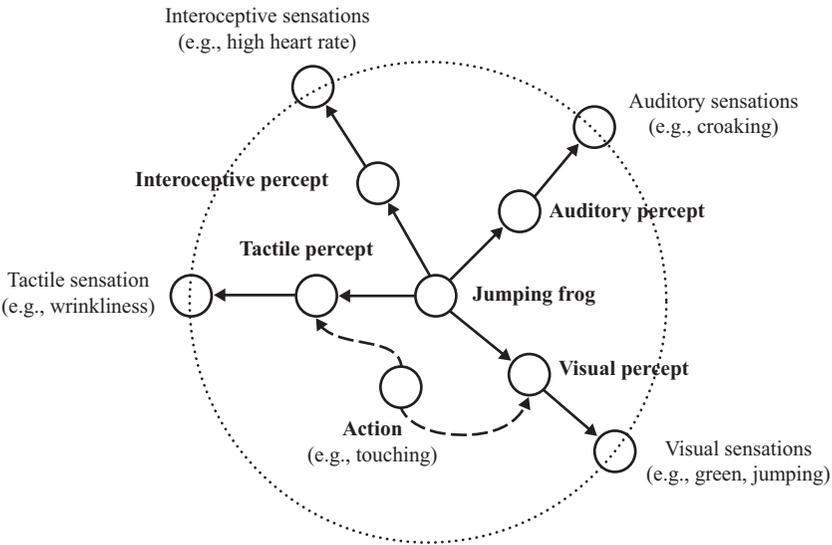
Figure 6.2
(Hierarchical) generative model for the concept of a jumping frog uses a simplified notation compared to chapter 4: nodes within the dotted circle correspond to hidden states, whereas nodes at the periphery correspond to sensory observations. Beliefs about hidden states, following inversion of the model, correspond to percepts that may be tied to a sensory modality (e.g., visual percept) or may be amodal (e.g., the jumping frog). Action contingencies are represented as dashed lines. Horizontal dependencies between hidden states in different modalities, as well as temporal dependencies between hidden states (as we saw in the dynamical generative models of chapter 4), are ignored for the sake of simplicity.

Box 6.1

Varieties of sensory modalities: Exteroceptive, proprioceptive, and interoceptive

In Active Inference, a conceptual distinction is often made between three kinds of sensory modalities: exteroceptive (e.g., vision and audition), proprioceptive (e.g., the sense of joint and limb positions), and interoceptive (e.g., the sense of the internal organs of the body, such as heart and stomach). In multimodal generative models, one can often factorize parts of the model that relate to different modalities; this permits representing that (for example) saccadic movements have visual but not auditory consequences.

Importantly, the same principles of Active Inference operate across all the modalities. For example, in the same way visual processing can be described as the inference about (hidden variables about) a perceptual scene, interoceptive processing can be described as the inference about (hidden variables that report) the internal state of the body. Furthermore, motor actions that change the perceptual scene and internally directed actions that change the interoceptive state can be described in a similar way. The former engages spinal reflexes that fulfill proprioceptive predictions, whereas the latter engages autonomic reflexes that fulfill interoceptive predictions. Such interoceptive processing supports allostasis and adaptive regulation, and its dysfunctions can have psychopathological consequences (Pezzulo 2013, Seth 2013, Pezzulo and Levin 2015, Seth and Friston 2016, Allen et al. 2019).

a single (multimodal or supramodal) hidden state at the center of the figure unfolds in a cascade of (unimodal) hidden states corresponding to percepts in different modalities (exteroceptive, proprioceptive, and interoceptive; see box 6.1) and ultimately causing sensations in the same modalities. This arrangement corresponds to casting the jumping frog concept as the *common cause* of multiple *sensory consequences* (e.g., something green and jumping in the visual domain; a croaking sound in the auditory domain), some of which can be action-contingent (e.g., the sight of something jumping may increase on touching it). The inversion of the generative model corresponds to a perceptual inference (e.g., the presence of a jumping frog) from its observed sensory consequences (e.g., the sight of something green and jumpy), and it integrates information across multiple modalities.

Once these variables of interest have been established, the next exercise is to write down the full generative model. One example is the simple generative model for frogs and apples in figure 2.1, which is fully specified by prior beliefs about hidden states and a (likelihood) mapping between

hidden states and observations and whose numerical values can be either specified by hand or learned from data (see 6.5.2).

Beyond this simple example, the elements that need to be specified are fully determined by the form of the selected generative model. For example, the model for discrete-time POMDP shown in figure 4.3 (top) requires specifying the **A**, **B**, **C**, **D**, and **E** matrices; continuous schemes use analogous (although less alphabetical) elements, which will be dealt with in chapter 8. But even in these more complex cases, the exercise is not so dissimilar from above: namely, specifying prior beliefs about the variables of interest (e.g., in discrete-time implementations, about hidden states at the first time step in the **D**-vector and about observations in the **C**-matrix) and their probabilistic mappings (e.g., likelihood mapping between hidden states and observations in the **A**-matrix). However, in some cases, it is useful to think about *factorizations* of the state-space of the generative model, which avoids considering every possible combination of variables if some are unnecessary. In chapter 7, we will discuss a biologically plausible example of factorization that occurs in perceptual processing between "what" and "where" streams (Ungerleider and Haxby 1994)—namely, between variables that represent object identities and locations, respectively, which can be treated independently in the model (hence simplifying it) as they are often invariant to one another.

Deciding which variables are of interest and the ways they are related or factorized in the model is often the most challenging—but also the most creative—part of model design. It is an exercise of translating our cognitive hypotheses into a mathematical form that supports Active Inference. How should we select the "right" variables? Ultimately, this is a question of specifying plausible alternatives and picking those that have the lowest free energy (cf. Bayesian model comparison). However, a practically useful perspective for most studies is that the generative model should be as similar as possible to how we believe data are generated. When appealing to Active Inference in the setting of cognitive psychology, this often means thinking about how experimental psychologists would go about generating the stimuli they present to their experimental participants. On formalizing these processes in terms of the requisite probability distributions, we arrive at a generative model whose free energy minimizing dynamics naturally lead to performance of the task in question.

Here, we can draw an analogy with most Bayesian (or ideal observer) models of perception, in which the models are designed to mimic (to a

Box 6.2
Priors and empirical behavior

Another perspective on the issue of selecting priors draws from a set of results known as the *complete class theorems* (Wald 1947, Daunizeau et al. 2010), which state that any statistical decision procedure (i.e., behavior) may be framed as Bayes optimal under the right set of prior beliefs. This means that if we are interested in explaining empirical behavior, our challenge is to identify the generative model (comprising prior beliefs) that would reproduce that behavior as simply as possible. In short, priors are a statement of a hypothesis about the system in question. If other prior beliefs would be plausible, this offers an opportunity to put this to empirical data through Bayesian model comparison. This also has implications for computational phenotyping in clinical populations. That there will always be a set of prior beliefs that render behavior Bayes optimal implies the key question—in understanding the computational deficits that give rise to psychiatric or neurological syndromes—is what these priors are. This idea is slightly counterintuitive at first. However, the complete class theorem means that asking whether a behavior is (Bayes) optimal is meaningless. The important question is, What are the prior beliefs that would make this optimal? In chapter 9, we will see how an appeal to free energy minimization based on our own beliefs as scientists offers a way to answer this question.

large extent) the structure of the task at hand, as in the example of recognizing a frog or an apple (chapter 2). This idea is sometimes equated with the *good regulator theorem* (Conant and Ashby 1970), which says that to regulate an environment effectively, a creature (whether biological or synthetic) must be a good model of that system. From the perspective of eco-niche construction, this is sometimes phrased in terms of the (statistical) fitness (Bruineberg et al. 2018) of a creature's model to its environment (and vice versa). However, this does not mean that an agent's generative model has to be identical to the generative process that actually generates data. For most practical applications, it can be simplified or different. We will return to this point later in this chapter (6.6).

6.5.2 Which Parts of the Generative Model Are Fixed, and What Is Learned?

Another design choice is deciding which parts of the generative model are fixed and which ones are updated over time as an effect of learning. In principle, Active Inference allows every part of the model—and even its structure—to be

updated (or learned) over time. This renders learning a design choice rather than something mandatory. In keeping with this, we will cover examples of Active Inference models that are completely designed by hand and examples in which some parts of the model (e.g., transition probabilities) remain fixed while others (e.g., likelihoods) are updated over time.

In Active Inference, learning is cast as an aspect of inference, as a free energy minimizing process. So far, we have described inference in terms of an *update of beliefs about states* of the generative model. In much the same way, we can describe learning as an *update of beliefs about parameters* of the generative model. For this, the generative model has to be endowed with *prior beliefs about parameters* of the distributions to be learned, where the specific parameters depend on probability distribution associated with each variable (e.g., *mean* and *variance* for a Gaussian distribution). These prior values are updated to form posterior beliefs whenever new data are encountered. As we will discuss in chapter 7, the algorithmic form of this update is the same as the update of state variables.

The fact that both inference and learning use the same kind of Bayesian belief updates may seem confusing during model design—partly because deciding what should be modeled as a state or a parameter is not always straightforward. However, when it comes to cognitive models, there is a clear difference between inference and learning. Inference describes (fast) changes of our beliefs about model states—for example, how we update our belief that there is an apple in front of us after observing something red. Learning describes (slow) changes of our beliefs about model parameters—for example, how we update our likelihood distribution to increase the value of the apples-red mapping after observing several occurrences of red apples. Beliefs about parameters typically vary much more slowly than those about states, and they may only be updated after states have been inferred. From a neurobiological perspective, it is appealing to map inference to neuronal dynamics and learning to synaptic plasticity. Furthermore, as we will discuss in chapter 7, holding probabilistic beliefs about model parameters induces novelty-seeking behaviors so that creatures may select the best data to learn the causal structure of their worlds. This suggests that endowing Active Inference models with the ability to learn their *parameters* (or even their *structure*; see chapter 7) is an effective way to study the behavioral dynamics of active learning and curiosity-based exploration.

Before concluding this section, it is worth noting that in this book we exemplify rather simple generative models that are defined using *tabular*

methods (e.g., with explicit matrices for priors and likelihoods) and that operate in small state-spaces. In comparison, much more sophisticated kinds of generative models—and associated learning schemes—are being developed in fields like machine learning, deep learning, and robotics, such as, for example, variational autoencoders (Kingma and Welling 2014), generative adversarial networks (Goodfellow et al. 2014), recursive cortical networks (George et al. 2017), and world models (Ha and Schmidhuber 2018). In principle, one could borrow any of these methods (and many others) to implement one or more parts of Active Inference models (e.g., likelihood or transition models). By leveraging the most up-to-date machine learning methods, it would be possible to scale up Active Inference to increasingly more challenging domains and applications; see, for example, Ueltzhöffer (2018) and Millidge (2019).

However, there are some important points to consider when designing Active Inference models that use sophisticated machine learning models, especially if one is interested in cognitive and neurobiological implications. One appeal of Active Inference is that it offers an integrative perspective on cognitive functions by assuming that (for example) perceptual inference, action planning, and learning all stem from the same free energy minimization process. This integrative power would be lost if (for example) one juxtaposed generative models that operate or learn independently from one another. Furthermore, the aforementioned machine learning methods correspond to process models that are distinct from Active Inference and have different cognitive and neurobiological interpretations. Finally, when using machine learning methods, some of the design choices discussed here (e.g., about the choice of model variables) may be skipped, as they are emergent properties of learning; however, they may be replaced by different design choices, about (for example) number of layers, parameters, and learning rates of a deep neural net. These design choices potentially have relevant cognitive and neurobiological implications, which are beyond the scope of what we address here.

6.6 Setting Up the Generative Process

In Active Inference, the *generative process* describes the dynamics of the world external to the Active Inference agent, which corresponds to the process that determines the agent's observations (see figure 6.1). It may seem bizarre to have postponed defining the generative process until after describing the

agent's generative model. After all, a modeler would have some task (and generative process) in mind from the beginning, so it would make perfect sense to revert this order and design the generative process before the generative model, especially in applications where the generative model has to be learned during situated interactions, as in gamelike or robotic settings (Ueltzhöffer 2018, Millidge 2019, Sancaktar et al. 2020).

The reason we postponed the design of the generative process is that, in many practical applications discussed in this book, we simply assume that the dynamics of the generative process are the same as, or very similar to, the generative model. In other words, we generally assume that the agent's generative model closely mimics the process that generates its observations. This is not the same as saying that the agent has perfect knowledge of the environment. Indeed, even if the agent knows the process that generates its observations, it may be uncertain about (for example) its initial state in the process, as was the case in the apple versus frog example. In the language of discrete-time Active Inference, one could design a model in which both the generative model and the generative process are characterized by the same **A**-matrix but in which the agent's belief about its initial state (**D**-vector), which is part of its generative model, is different from—or even inconsistent with—the true initial state of the generative process. One subtle thing to notice is that even if both the generative model and the generative process are characterized by the same **A**- and **B**-matrices, their semantics are different. The **A**-matrix of the generative process is an objective property of the environment (sometimes called a *measurement distribution* in Bayesian models), whereas the **A**-matrix of the generative model encodes an agent's subjective belief (called a *likelihood function* in Bayesian models).

Of course, except in the simplest cases, it is not mandatory that the generative model and generative process are the same. In practical implementations of Active Inference, one can always specify the generative process separately from the generative model, either using equations that differ from those of the generative model or using other methods, such as game simulators, which take actions as inputs and provide observations as outputs (Cullen et al. 2018), thereby following the usual action-perception loop implied by the Markov blanket of figure 6.1.

There are some philosophical implications of designing generative models that are similar or dissimilar from the generative process (Hohwy 2013; Clark 2015; Pezzulo, Donnarumma et al. 2017; Nave et al. 2020, Tschantz

et al. 2020). As discussed above, the *good regulator theorem* (Conant and Ashby 1970) says that an effective adaptive creature must *have* or *be* a good model of the system it regulates. However, this can be achieved in various ways. First, as discussed so far, the creature's generative model can mimic (at least to a great extent) the generative process. Models developed in this way may be called *explicit* or *environmental* models, given the resemblance between their internal states and the environment's external states. Second, the creature's generative model can be much more parsimonious than (and even significantly different from) the generative process, to the extent that it correctly manages those aspects of the environment that are useful to act adaptively in it and achieve the creature's goals. Models developed in this way may be called *sensorimotor* or *action oriented*, as they mostly encode action-observation (or sensorimotor) contingencies and their primary role is supporting goal-directed actions as opposed to providing an accurate description of the environment.

The difference between explicit and action-oriented models can be appreciated if we consider different ways one can model (for example) a rodent trying to escape from a maze in which some corridors are dead ends. An explicit generative model may resemble a cognitive map of the maze and provide a detailed characterization of external entities, such as specific locations, corridors, and dead ends. This model may permit the rodent to escape from the maze using map-based navigation. An action-oriented model may instead encode contingencies between whisker movements and touch sensations. This latter model would afford the selection of contextually appropriate strategies, such as moving forward (if no touch sensation is experienced or expected) or changing direction (in the opposite case)— eventually permitting the rodent to escape from the maze without explicitly representing locations, corridors, or dead ends. These two kinds of model prompt different philosophical interpretations of Active Inference, considering generative models as ways to either reconstruct the external environment (explicit) or afford accurate action control (action oriented).

Finally, as discussed in the field of *morphological computation* (Pfeifer and Bongard 2006), some aspects of a creature's or a robot's control can be outsourced to the body and hence do not need to be encoded in its generative model. One example is the *passive dynamic walker*: a physical object resembling a human body, composed of two "legs" and two "arms," which is able to walk an incline with no sensors, motors, or controllers (Collins et al.

2016). This example implies that at least some aspects of locomotion (or other abilities) can be achieved with body mechanics that are carefully tuned to exploit environmental contingencies (e.g., an appropriate body weight or size to walk without slipping); therefore, these contingencies do not need to be encoded in the creature's generative model. This suggests an alternative way to design Active Inference agents (and their bodies) that *are*—as opposed to *have*—good models of their environment. Yet all the ways to design Active Inference models are not mutually alternative but can be appropriately combined, depending on the problem of interest.

6.7 Simulating, Visualizing, Analyzing, and Fitting Data Using Active Inference

In most practical applications, once the generative model and generative process have been defined, one only needs to use the standard procedure of Active Inference—the descent of the active and internal states on the free energy functional associated with the model—to obtain numerical results. Arguably, modelers' goals are to simulate, visualize, analyze, and fit data (e.g., conduct model-based data analysis). Standard routines for Active Inference that provide support for all these functions are freely available (https://www.fil.ion.ucl.ac.uk/spm/); an annotated example of using these routines is provided in appendix C.

Although in most cases Active Inference procedures function off-the-shelf, in some practical applications one may consider specific fine-tunings or changes. For example, specifying the temporal depth of planning defines how many future states are considered during expected free energy computations. Setting up a limited temporal depth, along with other approximations to exhaustive search such as sampling (Fountas et al. 2020), may be useful in practical applications of Active Inference in large state-spaces.

Another example of adapting the standard functioning of Active Inference is the selective removal of parts of the expected free energy equation. This ablation may be useful to compare standard Active Inference (that uses expected free energy) with reduced versions, in which some parts of the expected free energy are suppressed to render them formally analogous to (for example) KL control or utility maximization systems (Friston, Rigoli et al. 2015). Furthermore, one can also augment Active Inference models with additional mechanisms, such as habitual learning (Friston, FitzGerald

et al. 2016) or learning rate modulation (Sales et al. 2019), with the caveat that maintaining the normative character of Active Inference would require casting these additional mechanisms in terms of free energy minimization.

Finally, other fine-tunings or changes to Active Inference may be useful to characterize disorders of inference and psychopathological conditions—for example, to explore the behavioral and neuronal consequences of endowing a creature's generative model with excessively strong (or weak) priors via excessively high (or low) levels of neuromodulators. We will provide some examples of Active Inference models that are relevant for psychopathology in chapter 9.

6.8 Summary

In this chapter, we have outlined the most important design choices that must be made in setting up an Active Inference model. We provided a recipe in four steps and some guidelines to address the usual challenges that model designers face. Of course, it is not necessary to follow the recipe in a rigid manner. Some steps can be inverted (e.g., design the generative process before the generative model) or combined. But in general, these steps are all required. This sets up the remainder of this book, which puts these ideas into practice through a series of illustrative examples designed to showcase the theoretical principles presented in the first half of the book. In everything that follows, the only differences among the examples rest on the design choices we have highlighted here. Part 2 illustrates systems with different boundaries, with discrete or continuous dynamics at different timescales, for which the choice of prior beliefs is fundamental in reproducing behavior across many different domains—but all implementing the same Active Inference.