

## Chapter 22

### A Bayesian conversation

Nick Chater, Thomas L. Griffiths &  
Joshua B. Tenenbaum

What are the prospects for Bayesian cognitive science? Is it too specific, making claims about the mind that are already known to be false? Or too general, providing a framework so capacious that any empirical data can fit comfortably within it without danger of falsification? How does the Bayesian approach relate to other work in cognitive science? In this chapter, we discuss these issues in the form of a conversation between a skeptic (S) and a Bayesian (B) cognitive scientist.

To do so, we imagine our hypothetical skeptic having concerns arising from several different perspectives, such as traditional cognitive psychology and connectionist cognitive modeling. We draw upon questions from actual skeptics (McClelland et al., 2010; Jones & Love, 2011; Bowers & Davis, 2012) and our responses to them (Griffiths et al., 2010; Chater et al., 2011; Griffiths, Chater, Norris, & Pouget, 2012). In some cases, our responses serve to clarify potential misunderstandings, or weaken what may be perceived as overly strong or confident claims. In other cases, the challenges raise important open questions for the Bayesian approach that future work should seek to address. We hope the ideas below will stimulate further debate and research from readers of this book to push forward the development of cognitive science, whether within the Bayesian framework or outside it.

So let the debate begin!

*S: The premise of this book is that cognitive science is an exercise in reverse engineering. Let's suppose that is right. Then we should be looking for inspiration for the cutting-edge engineering methods that are driving forward contemporary machine learning and AI. But surely many of the real breakthroughs in AI and machine learning aren't Bayesian at all. The biggest game-changer in the past decade or two in machine learning has been the rise of deep learning (Goodfellow et al., 2016). This is the technology that underpins google images, autonomous driving, some incredible recent results in natural language processing, and has helped computers finally beat the best human at Go and a huge range of other games (Brown et al., 2020; Silver et al., 2017). The Bayesian approach is philosophically well-justified and mathematically elegant, perhaps. But isn't it just blown away by the general-purpose power of deep learning from really large amounts of data?*

B: The results of deep learning are undoubtedly remarkable, and have led to incredibly useful practical applications. But these methods aren't succeeding by modeling human intelligence. Instead, they are solving specific, and practically important, problems by methods that appear entirely to side-step the need for human intelligence.

What is remarkable about people is their ability to learn rapidly, reason flexibly, and think creatively, about entirely novel problems. To do this, our brains are able to make wild inferential leaps from minimal amounts of data. By contrast, deep learning typically involves making comparatively small generalizations from extremely large amounts of data. Now of course deep learning is not just table look-up; it is more like a very sophisticated table look-up with extremely clever interpolation. This type of mechanism may very well be important in understanding many aspects of cognition, particularly concerning problems of pattern recognition that arise in and motor control. But surely much more is required to help engineer machines that learn like humans (Muggleton & Chater, 2021).

Lake et al. (2017) stress the importance of three key elements, which have all figured heavily in this book: the ability to build causal models of the external world rather than purely finding patterns in data (Chapters 4, 5 and 9), the capacity to learn and apply rich background theories of how the world works (including naïve physics and folk psychology, Chapters 14 and 15); and the ability to construct compositional representations (Chapters 17-20) and to learn to learn (Chapter 8).

In short, the response to your question is: yes, deep learning is terrific engineering, but it doesn't reverse-engineer human cognition (though it might be part of the story). It just doesn't solve the deep problems of human intelligence, and doesn't operate in a human-like way. Or at least not so far.

Consider a parallel with the body. Understanding the heart as a pump was a spectacularly successful piece of reverse engineering, which would have been impossible without the invention of pumps for moving

water. And the valves in the circulatory system would be hard to understand without a prior knowledge of the engineer’s conception of a valve. But the complex set of dynamically contracting and relaxing chambers that make up the heart work very differently from any pumps known in engineering. And the most efficient pumps in engineering certainly do not mimic the solution found by natural selection. Or, to take a more extreme example, the problem of reverse-engineering the biomechanics of human locomotion depends on very complex background engineering knowledge; but it does not depend on the technology of internal combustion engines, gears, or wheels.

*S: That all sounds plausible enough—but are you sure that deep learning isn’t more human-like than you think? Look at the incredible results obtained by GPT-3, GPT-4 and other large language models which we discussed briefly in Chapter 16—ranging from generating pastiche chunks of period novels, “fake” but strangely plausible philosophical conversations on consciousness, decent-looking computer code, answers to questions on a huge variety of topics, and much more. Who is to say what is going on deep inside the one hundred and seventy five billion parameters in GPT-3’s neural network? Maybe GPT-3, GPT-4 or one of their many rival models have already developed causal models of the world, figured out rudimentary naïve physics and psychology, and created compositional representations from scratch. Or maybe some future even more powerful deep neural network model will soon do so? Perhaps the Bayesian approach to cognition, and the importance of structured representations, and so on, is perfectly fine, but really a distraction. The real action is going to be building the ever-more powerful learning algorithms, trained on ever-richer sources of data (perhaps not just words and images but something closer to the rich sensory-motor input available to a human child) that will explain how all this complexity can be created from scratch.*

B: Well, even if that turns out to be right, then the Bayesian story will be still be important: We will still want to understand how giant neural network models are working, just as we will want to understand the operation of the human brain. But the evidence so far strongly suggests that at least early large language models using deep learning, like GPT-3, work in profoundly non-human-like ways. Remember how in Chapter 16 we saw that GPT-3 seems to do language processing by something like a highly sophisticated table look-up, with incredibly clever cut-and-paste interpolation, rather than having the remotest conception that language has meaning, is used in communication, or that there is an external world to be communicated *about*. GPT-3 knows a lot of “facts” in domains in which its training corpus (a large fraction of the entire internet) contains relevant sentences—so it knows that spiders have eight eyes even though many people don’t. But, as we saw, it has no idea that a foot has no eyes, because the internet is not full of discussions of this (rather nonsensical) topic. A human, with even the most rudimentary knowledge of biology, knows that feet have no eyes; but GPT-3 does not have this rudimentary understanding—indeed, this might suggest that GPT-3, at least, has no real understanding at all. Or, to put the point more neutrally, the understanding of such models is at least hotly contested (Mitchell & Krakauer, 2023).

*S: But the new generation of deep learning systems, or the generations after these, might spontaneously develop such understanding. GPT-4’s flexibility and intelligence may not merely be papering over the cracks of earlier models, but learning to work in a fundamentally different, and more human, way. And who knows what will be possible with large neural networks in a decade, or perhaps a century.*

B: Time will, of course, tell. For now, there are at least good reasons to believe that the apparent “sparks of general intelligence” that even the most sophisticated large language models exhibit may have a more prosaic statistical basis, and may fail in ways that seem unnatural from the perspective of human cognition (McCoy, Yao, Friedman, Hardy, & Griffiths, 2023). Indeed, it may well turn out that relying on deep learning or similar approaches to solve the fundamental problems of cognition will turn out to be like betting that the next generation of automotive technology will miraculously explain the details of human biomechanics. For reverse engineering to succeed, we can’t just rely on finding good solutions to engineering problems. We need to find engineering solutions to the actual challenges faced by human intelligence (Griffiths, 2020); and we need, also, to check that any successful system is solving these

problems in a human-like way.

Suppose, though, that future generations of neural network model do seem not merely to be emulating the results of human intelligence, but can build a real understanding of how the world, and other humans, work—or at least can build as much of an understanding as we humans have. Then there is still the question of how these results are achieved. And the Bayesian reverse-engineering approach may turn out to be crucial to answering this question. Just specifying the learning algorithm, the architecture of the network and its sources of data isn't enough to give us much insight into how such networks are working. We still wouldn't know what knowledge it represents, and in what form. In fact, the challenge of figuring out how a giant artificial network works is not so different from the challenge of understanding the biological neural network that is the human brain. Of course, the artificial neural network is, of course, in principle far more “transparent”—we can see its activity in minute detail, rather relying on the relatively crude results of brain imagining. But these detail may not help much anyway, at least without some theory of the possible computations that the system is carrying out—and to make this possible we will most likely need the Bayesian analysis after all.

*S: Well, I can see we're not going to agree on this—and perhaps that's OK: having different research strategies running in parallel is usually a good approach in science, after all. But all this discussion of solving problems in a human-like way seems a double-edged sword for the Bayesian cognitive scientist. After all, Bayesian cognitive science sees the brain as an exquisite probabilistic reasoning machine. But surely this isn't a human-like model at all, but an assumption of hyper-rationality—what Gigerenzer and Goldstein (1996) have termed the “Laplacian demon.”*

*Doesn't this run in the face of developments in rational choice explanation in the social sciences? Surely, the whole point of behavioral economics, for example, is that such rational idealizations are incorrect—or at least incomplete; and these disciplines look to the cognitive sciences to build more realistic, and hence less hyperrational models of behavior. This viewpoint seems to fit, too, with many and diverse traditions in psychology, the cognitive sciences and artificial intelligence, which sees intelligence as more like a flexible “bag of tricks” rather than the result of fully rational calculation (Agre & Chapman, 1987; Brooks, 1991; Gigerenzer & Todd, 1999; Ramachandran, 1990).*

B: But the Bayesian approach does not require that people necessarily carry out enormously complex Bayesian calculations. The Bayesian approach may, in some instances, specify the “right” pattern of inference or course of action; but the agent need not necessarily recapitulate this calculation in order to justify its thoughts and behavior, any more than the bird needs a knowledge of fluid dynamics in order to “justify” the shape of its wings (Marr, 1982). Instead, the cognitive system may come upon the right solution by evolution; through trial-and-error learning; or by some other means.

One of the striking contributions of approaches in cognitive science and artificial intelligence based on simple heuristics is to show that, in some environments, such approaches can be surprisingly successful. But why? A Bayesian rational explanation is surely required—spelling out an optimal solution, given the structure of typical environments, and showing that simple heuristics often come close to this optimum. It seems likely that, where such short-cuts are available, the cognitive system may use them—and a Bayesian analysis is helpful in identifying those situations.

Yet the “bag of tricks” perspective is fundamentally incomplete as a reverse-engineering account, in at least two ways. A complete solution would indicate, first, how the cognitive system knows which heuristic to apply in which circumstances, and second, and more importantly, how the cognitive system adjusts these heuristics in the light of new information. In a nutshell, an “adaptive toolbox” or “bag of tricks” model of cognition may be appropriate if there is a well-defined and stable “bag of problems” to which the tricks naturally correspond. But if the agent has to respond deftly and appropriately to a continually changing set of problems, which is surely the case with human cognition, then the agent needs to be guided by some principles concerning how to think or act. If these principles are to work effectively, then they must, presumably, have some justification—and hence we are forced back to, at

least some approximation to, a rational approach (indeed, the rational solutions to these problems are discussed at length in Chapter 13).

An analogy may be helpful. Given a fixed set of arithmetic problems, a good strategy may be simply to store the answers to each by rote, so that the questions can be answered rapidly without calculation (Logan, 1988). But if the set of problems is broad and unpredictable; and the scope of those problems is continually liable to shift (e.g., if negative numbers, fractions, reals or imaginary numbers are successively introduced), then it is vital to understand the general principles of arithmetic.

*S: But what about computational complexity limitations? Surely we've known for a long time that Bayesian calculations cannot possibly scale up to deal with real-world inference.*

B: As we have just noted, complex Bayesian calculations may sometimes end up explaining why the cognitive system sometimes sticks to simple heuristics—for which no complexity issues arise. But, in any case, in principle complexity results can be misleading, as they typically consider inference over arbitrary probability distributions. The magic of graphical models is that they provide a way of factorizing probability distributions into a particularly workable form—and a form which lends itself to parallel implementation: each node can be viewed as a processor, which makes calculations entirely locally, depending on its own inputs; but the resulting collective behavior may correspond to a globally “good” solution. More precisely, the computational cost of naïve probabilistic inference is exponential in the number of variables involved, but inference in a Bayesian network is exponential in the size of the largest clique in the underlying graph (Cooper, 1990). Performing exact inference in such a network may thus, nonetheless, be costly—but then numerical methods, such as Markov Chain Monte Carlo (Gilks et al., 1996) and other approximation methods that we met in Chapter 6, may allow good approximate solutions.

But, of course, issues of computational tractability are hugely important, from the point of view both of engineering and reverse engineering. Cognitive processes must run fast on the slow parallel hardware of the brain (Feldman & Ballard, 1982). But over the past few decades there has been great progress in finding ways in which seemingly intractable probabilistic problems can be approximated (see Chapters 4, 5, and 6). We suggest that these, and future, developments concerning tractable, approximate, Bayesian inference will be an important source of hypotheses in reverse engineering how the brain deals effectively with what may appear to be unmanageably severe computational challenges.

*S: But doesn't more than a half-century of research in judgment and decision making tell us quite the opposite? Rather than finding elegant approximations to Bayesian calculations, people seem to deviate drastically from them, even for the very simplest problems. Endless experiments have found that human probabilistic reasoning seems to be riddled with systematic errors and biases. People persistently have the wrong qualitative intuitions too: the conjunction fallacy, ignoring base-rates, the gambler's fallacy, Simpson's paradox, the Monty Hall problem, and many more surely make it clear that the human mind is not a Bayesian. Surely this is the elephant in the room for Bayesian cognitive science. It is completely incompatible with the ubiquitous inadequacies of human probabilistic reasoning.*

B: Yes, I'm glad you raised this! It is so easy to think: “I'm sure I've heard that Kahneman and Tversky have shown people aren't Bayesian. So this whole Bayesian cognitive science movement has got to be on the wrong track.”

But this would be a serious mistake! There are several points to make, so let's take them one by one. The first point is that the Bayesian cognitive models considered in this book, and similar models in computational neuroscience, computational models of perception and motor control, have absolutely no direct implications for human probability judgments.

Remember that cognitive models of all kinds involve complex mathematics. But the mathematics underpin the operation of the models. It would be a complete misunderstanding to imagine that such models imply that people have to understand and be proficient in the relevant branch of mathematics. This would be a mistake akin to imagining that the deep reinforcement learning can't possibly explain

how rats learn, because rats don't know even know basic algebra and calculus, let alone how to back-propagate a gradient; or that the rabbit retina can't be convolving the image with its receptive fields, because rabbits haven't learned about convolution; or, for that matter, that complex biochemistry can't underpin your own digestion unless you have a deep understanding of biochemistry. Indeed, Marr (1982) made this point forty years ago, noting that a bird does not need to know the principles of aerodynamics that make flight possible. Likewise, Bayesian models of cognition remain a useful tool even if we seek to understand a mind that is utterly ignorant of probability theory. So the elephant in the room turns out not to loom quite so large after all. The worry is primarily based on a conceptual confusion.

But that's not quite the end of the story. As we've seen throughout this book, the Bayesian computations involved in any cognitively interesting problem tend to be far too complex to solve precisely. So in practice Bayesian models, in cognitive science just as in statistics and machine learning, generally use approximation methods, of various types. It would particularly neat if it turned out that the biases and errors people fall into when they are reasoning about probabilities turn out to be side-effects of one of this process of approximation. Now, of course, a lot of the probabilistic reasoning tasks that people are given in laboratory experiments do have simple analytic solutions—so no approximation methods are required. But, of course, the probabilistic machinery of the brain is not adapted to simple verbal or numerical probability problems; so it is likely that the approximation algorithms used to solve complex problems will be used here too.

As we saw in Chapter 11 (see also, e.g., Dasgupta et al., 2017; Sanborn & Chater, 2016), one approach assumes that people make probability estimates through sampling, possibly with some process of correction for sample size (Zhu et al., 2020). Then, as we saw, a variety of biases will occur, potentially due to biases concerning where sampling begins, the fact that samples of autocorrelated rather than independent, and so on. And some of these biases map neatly onto observed probabilistic reasoning “biases” including conservatism, anchoring, representativeness, and so on (for a survey see Chater et al., 2020). So it is possible that understanding the mind as approximate Bayesian inference may provide an elegant and unifying explanation of apparently disparate quirks in human probabilistic reasoning. Thus, the frailties of human probabilistic reasoning may fit neatly within a Bayesian framework, rather than being a counterexample to it.

This viewpoint is part of a broader perspective on Bayesian cognitive science: that empirical predictions need not directly be drawn from “pure” Bayesian analysis of the probabilistic problems faced by the brain, but rather through understanding how these problems are approximated, given the computational machinery of the brain. This type of resource-rationality perspective can be developed in a number of ways (Griffiths et al., 2015; Lieder & Griffiths, 2020). Sticking with sampling models, one approach is to assume that, if computationally expensive, samples will be “rationally” biased to where they are likely to be most valuable (“utility-based sampling”; see Chapter 13). This will lead to an over-representation of rare but significant events, providing an explanation for the excessive influence of such events (e.g., plane crashes, terrorist attack) on probability judgments in a wide range of contexts (Lieder et al., 2018). Relatedly, resource-bounded rational models can provide a basis for rank-based models of sampling (Stewart et al., 2006), on the assumption that psycho-economic scales are unavoidably noisy (Bhui & Gershman, 2018).

*S: But doesn't it just still seem a bit weird that Bayesian cognitive scientists are postulating that the brain solves really hard probability problems easily (involving vision, language, categorization, and so on) and yet falls down completely on trivial probability problems?*

B: Really it shouldn't sound weird at all! Consider vision. Suppose that it turns out that human, and perhaps more broadly animal, vision, can be viewed as involving (suitably approximated) Bayesian inference over images, represented in a hierarchy of levels of representation. An agent with such a visual system would not thereby be expected to be able to engage in general probabilistic reasoning, over verbally stated problems—indeed, non-human species would inevitably lack this ability.

*S: Well that just brings me to another problem. Probability theory is all about quantifying uncertainty in numerical form—and making calculations that link different numerical uncertainties with each other. But most of the time people don't think about numerical probability at all—instead, we mostly reason about how the world works, and what will, might, or certainly won't, happen in purely qualitative terms. The brain is like a lawyer arguing that the defendant is, or is not, guilty—and our conclusion is determined by the qualitative structure of the arguments one way or the other.<sup>1</sup> Lawyers don't talk about probabilities of guilt; and jurors and judges wouldn't like it if they did. That's not the way mind works. Of course, in the last few centuries, human have developed a mathematics for dealing with numerical probability—but the very recency of probability theory, and the trouble each new generation of students has figuring out how to use it, surely testifies to how cognitively unnatural it is.*

B: I couldn't agree more. Yes, but it is a mistake to think that the Bayesian approach to cognition, or indeed inference in general, is about probabilities represented as numbers. It's really much better viewed as a set of tools for thinking about the structure of reasoning, and specifically reasoning about uncertainty.

Think about graphical models, which encode dependence relations between propositions (and, for causal graphical models, causal links which have implications about counterfactuals; see Chapter 4). The structure of models is really where the action is: it is this structure that underpins qualitative patterns of reasoning, showing which propositions depend on which others, what causes what, and so on.

Commonsense reasoning is incredibly rich and flexible, and dependent on rich background understanding—which, we've noted, is naturally modeled within a Bayesian framework. Thus, seeing the window smashed, and footprints in the flower bed below, the reasoner may use rich causal and cultural knowledge to infer that a burglary has taken place. Given the additional knowledge that a camera crew is nearby, the most likely hypothesis readily switches to the assumption that no real crime has been committed, and that instead a murder mystery is being filmed. Still further information, e.g., that the camera crew is carefully hidden and using long-range lenses, might flip the reasoner's most probable interpretation back again: the camera crew are police who have been tipped off about the burglary. Such reasoning involves drawing on extremely rich knowledge of the world; indeed, the knowledge that may be engaged seems entirely open-ended.

The engineering project of attempting to capture such inferences, although initially framed in terms of logical reasoning, is now most commonly addressed in artificial intelligence in probabilistic terms. Graphical models have been particularly fruitful ways of representing people's knowledge of the world (Pearl, 1988); and, in artificial intelligence, there have been important steps in understanding how people can reason about causal relationships, using probabilistic methods (Pearl, 2000). The Bayesian perspective is a powerful framework for such analysis—so, in the above example, a boost to the probability of one hypothesis (“fictional crime reconstruction”), reduces the probability assigned to alternative hypothesis (“genuine crime committed”), a pattern of inference known variously as causal discounting (in psychology; Einhorn & Hogarth, 1986; Kelley, 1987) or explaining away; (in artificial intelligence, Pearl, 1988), that we have encountered in various places in this book (particularly Chapter 4). Note, moreover, that patterns of commonsense reasoning, both in adult and infant cognition, is often viewed as analogous too, and perhaps continuous with, scientific reasoning—where again the Bayesian approach is now a dominant mode of explanation (Horwich, 1982; Howson & Urbach, 1993; Bovens & Hartmann, 2004) (though see Gelman & Shalizi, 2013, for an important dissenting perspective). Moreover, many laboratory studies of human reasoning, once interpreted as indicating that people violate logical patterns of reasoning, can be reinterpreted as indicating that patterns of human reasoning neatly accord with Bayesian principles (e.g., Hahn & Oaksford, 2007; Oaksford & Chater, 1994, 1998a, 2007, 2020), given appropriate assumptions about the background knowledge people possess about the domain about which they are reasoning.

---

<sup>1</sup>In fact, a good case can be made that verbal arguments between people may even underpin reasoning and argumentation within an individual mind (Mercier & Sperber, 2011).

*S: So let me take a different tack. Even from the earliest developments in probability theory, it was assumed that the calculus of probability should help to clarify the nature of everyday human reasoning about uncertain events. Indeed, Bernoulli's book, *The Art of Conjecture*, explicitly aims both to characterise and to improve human reasoning. And after all, the subjective interpretation of probability as "degree of belief," which is the starting point of the Bayesian approach, surely should have some relationship with the intuitive notion of belief – perhaps the most fundamental notion in folk psychology. So this raises two issues. First, to what extent is the Bayesian approach them attached to the folk psychological conception of the mind is driven by beliefs and desires; And surely, since Nisbett and Wilson (1977), and through the rich tradition in social and cognitive psychology since Johansson, Hall, Sikstrom, and Olsson (2005), it has become apparent that our explanations of our thoughts and behavior in terms of beliefs and desires are, at best highly suspect, and at worst complete fictions (Chater, 2018; Churchland, 1981). But if it turns out that this folk psychological perspective is unworkable, where does this leave the Bayesian approach?*

*Second, if we do want to maintain a direct connection between the concept of belief in the Bayesian approach to cognition and the folk psychological notion of belief, is it really possible to do so? That is, can credibly claim that probability theory is rich enough to describe patterns of everyday inference? Or the entire machinery of logic and probability? Surely the philosophy of the later-period Wittgenstein (Wittgenstein, 1953), the apparently unruly nature of real common-sense and scientific inference (Feyerabend, 1993; Lakatos, 1970), the frame problem in artificial intelligence (Ford & Pylyshyn, 1996; McCarthy & Hayes, 1981), and many related lines of thought should undermine the idea that human reasoning can be reconstructed in a formal framework of any kind.*

B: These are deep issues, and the question of the relationship between "degree of belief" and the everyday notion of belief is, indeed, of fundamental importance. The mathematical and computational resources of the Bayesian approach are, of course, still available even if one rejects any such relationship. So, for example, Bayesian models in computer vision and computational linguistics make reference to a kind of low-level linguistic additional features which have absolutely no counterparts in "folk psychology," and hence are not the kind of thing about which we are typically assumes to have beliefs, in the full-blown sense. The rules of belief propagation in a Bayesian network can be applied, after all, irrespective of the interpretation of signs to the nodes of the network, if any. So the notion of belief for the application of formal Bayesian methods is a very reduced one. So, if the folk psychological enterprise were to turn out to be no more scientifically respectable than folk physics or folk biology as some have argued (Chater, 2018; Churchland, 1981; Rosenberg, 2019), Bayesian cognitive science would in no way be threatened.

Still, though, we see the potential connection with commonsense beliefs, typically expressed in natural language, is one of the most important selling points for the Bayesian approach: forging close links between cognitive modelling and everyday notions of belief, desire, reasoning, and knowledge. In many areas of psychology, the dominant style of explanation is not computational, but folk psychological: social psychologists explain people's behavior in terms of consonance or dissonance between their beliefs or attitudes; clinical psychologists aim to explore the abnormal belief systems formed by people with mental disorders. But the same applies within the cognitive science of high-level thought, which is similarly steeped in talk of categories, inferences, and knowledge, notions of transparently folk psychological origin. Equally, developmental psychologists characterize many aspects of cognitive development as involving the acquisition of increasingly rich beliefs and concepts, perhaps organized into theories of the world. This has allowed Bayesian models to capture a wide range of phenomena in cognitive, social and developmental psychology, which are usually conceived of in folk psychological terms.

This is, moreover, an important virtue of the Bayesian approach for another reason: that models of human intelligence need to capture the human ability to fluently generate and understand language—and to relate this understanding to our internal models of the world. If our internal models happen to be interpretable in intuitively meaningful terms (e.g., as capturing claims about what causes what), then these internal models can, in principle, be reported by the cognitive system, and modified in response to linguistic input from others.

The contrast with neural network models, and earlier and more primitive associative models of learning is interesting. A series of now rather little-known studies (reviewed in Brewer, 1974) showed that associative learning turns out to be surprisingly easily manipulated by verbal information. A typical study might train people to expect an electric shock in response to a cue, and measured the strength of this association implicitly by measuring skin conductance after the stimulus was presented. They then explicitly told people that the contingency would no longer hold (e.g., by saying something like “I’m turning the shock machine off now”). People’s skin conductance responses declined sharply; and declined in proportion to how much the participants reported that they believed the experimenter’s reassurance. More recent work similarly suggests that associations are formed, and modified, by general processes of knowledge revision, rather than by a distinct, mechanistic associative mechanism (Mitchell, De Houwer, & Lovibond, 2009). This makes sense if people are constructing and training graphical models, with links that they can abruptly disable when they hear that a previously active causal link has been severed. But it’s hard to understand how this can work if, for example, knowledge of causal connections is distributed in an impenetrable way throughout the weights in a vast deep neural network.

A probabilistic “language of thought” perspective (Chapters 16-18, Piantadosi & Jacobs, 2016) takes this approach a step further. If the kinds of symbolic knowledge expressed in natural language can be captured in a rich internal symbolic language, then the interface between that knowledge and natural language becomes potentially more transparent. But there need be no necessary commitment that the “sentences” in such a language refer about the objects and categories of natural language (see, e.g., Dennett, 1978). More broadly, the less direct the connection between internal representations and natural language, the more work will be required in building a computational theory of communication (see, e.g., Christiansen & Chater, 2022; Fodor, 1975; Goodman & Frank, 2016, for contrasting perspectives).

*S: But even suppose this is right—what about my second concern: that probability theory just isn’t up to the job of capturing the complexity everyday inference*

B: Well, probability theory is only part of the story – although we believe it to be an important part. What probability theory does is to help map from a current knowledge state, to an appropriate future knowledge state, in the light of new data that come to system has received. But the challenge of characterising this knowledge, i.e., describing the content of the naive theories of the natural and social world which guide our behavior, and representing these in some formal machinery, remains an enormous challenge.

*S: That may turn out to be a considerable understatement! The project of developing classical symbolic artificial intelligence foundered on this very problem, after all. Consider the notorious frame problem (McCarthy & Hayes, 1969; Ford & Pylyshyn, 1996), which is roughly that, whenever an agent updates a belief, or takes an action, it cannot be sure which of its other beliefs it needs to update. In short, the problem is that each individual belief can have the implications throughout the entire network of beliefs, and these implications can be very far-reaching but also very unexpected. So I might, for example, be operating with a particular intuitive model of the functioning of, say, the appliances in my house, but a single unexpected observation (for example seeing the streetlight outside my house snap “off” without warning) might lead me to suppose that there may be a power cut, and hence that my freezer or television will cease to function. Of course, one can always in large ones model to incorporate such possibilities.*

*So I may have a model of the functioning of my appliances which includes the possibility of power failure and its consequences (the contents of my freezer begin to thaw, and my television screen goes blank). Moreover, one might suppose that the inference from the streetlight going out, to there being a power failure can itself be captured in probabilistic terms. But the problem is not that each extra complication cannot be handled piecemeal, but rather that the open-ended character of the world knowledge requires that the process of extending and elaborating the probabilistic model seems to have no obvious end. Noticing the distant glow of light in another room, I might conclude that the power supply to the house must be intact after all; recalling that one of my lamps is powered by a battery, I may consider whether to retract this inference; noticing a repair van pulling up next to the streetlight might lead me to suspect that the bulb*

*has blown, and hence that there is no reason to suppose that the power supply is not intact; noticing that the man emerging from the van lives next door may undercut to this conclusion, because I suspect that he has not arrived to repair the bulb but has simply returned home. Each of these inferential steps can, to be sure, be reconstructed, perhaps quite neatly, in a Bayesian framework; but surely the implication is that our general intuitive knowledge of the world cannot be reconstructed as a set of stable and well-defined probabilistic models, such as those described in this book.*

*But this seems hopelessly implausible. Indeed, Leonard Savage, one of the key pioneers of the foundations of Bayesian statistics, explicitly distinguished between “small worlds” which are simple enough for a person to reasonably expected to have consistent degrees of belief (and hence a probabilistic model capturing these degrees of belief can be constructed), and “large worlds” where we will inevitably be inconsistent, and no probabilistic model is possible (Savage, 1972). Savage thought the Bayesian approach only appropriate for small worlds; yet Bayesian cognitive science seems blithely, and implausibly, to assume that the same approach applies quite generally (Binmore, 2008).*

*At best, the Bayesian seems to be left attempting to make sense of one, gigantically complex and impossibly unwieldy probabilistic model of all of world knowledge, capturing the entire “web of belief” (Quine, 1960; Quine & Ullian, 1978). At worst, we might see the probabilistic approach to capturing world knowledge as the sinking into the same quicksand that devoured symbolic approaches to knowledge representation in artificial intelligence in the 1970s and 80s.*

B: These sorts of problems certainly raise serious challenges for cognitive science. But they certainly don't bear against the Bayesian approach specifically. After all, according to just about any theory, the brain is somehow able to construct a rich and highly flexible model of the natural and social world. But the natural and social world is of course the enormous complexity, and it does not come divided into neatly pre-sealed units, each governed by its own set of entirely independent principles. The formal and conceptual tools required to capture such structure will be many and varied; and this must surely be embodied, to some extent a least, in the complexity of our mental representations of this information. And the inference is that we draw, will be largely dependent on the contents of these mental representations, and only secondarily on general inference principles such as the laws of probability. One of the central challenges the artificial intelligence is to understand the nature of such knowledge.

Indeed, as many influential theoretical models of cognitive development suppose, the child (and indeed the adult) should be thought of as engaging in an activity analogous to science (e.g., Gopnik et al., 1999), then the contribution of the Bayesian approach to understanding cognition is analogous to the contribution of the Bayesian movement in the philosophy of science – i.e., helping to explain general patterns of reasoning. Only indirectly can it assist with the problem of characterizing the content of the intuitive “science” that underlies cognition.

The problem of understanding human knowledge is, therefore, not specifically a challenge to the Bayesian approach, but rather a challenge for the cognitive sciences in general. Probability theory therefore no more promises to solve problems of knowledge representation, than it promises to solve the problems of the natural and social sciences. What the Bayesian approach does give us, however, is a framework for helping to draw certain types of inferences from such knowledge – and perhaps some tools for helping to represent limited aspects of this knowledge. But the vast bulk of the problem of understanding human knowledge, and the inferences that can be drawn from it, concerns the nature of their knowledge itself.

These disclaimers aside, it may be the Bayesian methods may be helpful in characterizing important aspects of human knowledge. We have particularly stressed the importance of defining probabilistic knowledge and reasoning over rich, structured representations, including graphs, grammars, logics and programming languages. Related developments in the probabilistic representation of knowledge may also prove to be important. For example, Pearl (2000) and Spirtes et al. (1993) have made important steps towards in the understanding of reasoning about causality, and in particular to understanding the

difference between an observation and an intervention, working within a probabilistic framework, some aspects of which may be directly relevant to cognitive science (see also Gerstenberg et al., 2021; Glymour, 2001; Sloman, 2005).

Bayesian cognitive science, moreover, provides deep general ideas about how knowledge can be acquired, represented and applied, irrespective of topic. Ideas such as graphical models, belief propagation, Markov Chain Monte Carlo, and so on, may prove to be fruitful metaphor is understanding cognition. The Bayesian framework may also be essential in order to capture the general inferential principles by which different semiautonomous domains of knowledge can be “plugged together,” into a functioning reasoning system.

*S: In a well-known article, Gould and Lewontin (1979) point out one of the potential pitfalls of any style of explanation based on optimality. We may inadvertently begin to follow Voltaire’s Dr Pangloss to reframe the domain of study so that it becomes the “best of all possible worlds.” Gould and Lewontin (1979) were concerned about possible excesses in adaptationist explanation in biology—that the biologist may start to see all and every feature of a living organism as exquisitely adapted to some environmental challenge. They argued that, by contrast, many aspects of the structure of an organism are not adaptations at all—and arise as side-effects of structures which may be adaptations. Surely the Bayesian cognitive scientist falls into precisely this trap? Almost any aspect of thought and behavior can, with sufficient ingenuity, be interpreted as arising from an optimal Bayesian calculation, if we can freely vary assumptions about the goals and knowledge of the agent. As with adaptationist explanation in biology, any apparent counterexample (i.e., deviation from optimal design or behavior) are surely all too easy to explain away—leaving the Bayesian proposal unfalsifiable and empirically empty.*

B: This is a real concern—but no more for Bayesian explanation than for any other (Dennett, 1983). Any style of theorizing much be judged by the standard criteria of scientific explanation—not mere fit with the data, but, for example, simplicity, and successful prediction. And any scientific hypothesis can be saved to fit the data by suitable *ad hoc* adjustments. Thus, as Putnam (1974) notes, Newton’s Laws could be saved, even if we observed square planetary orbits, if suitable additional forces, of unknown origin, are assumed to be operating, in addition to gravity (see Stanford, 2017). So the Bayesian cognitive scientist is in no better, and no worse, position than advocates of any other approach, with respect to being able to “save the theory,” come what may; and, as with other approaches, such attempts will ultimately be fruitless, because theory will become increasingly complex, and less predictively successful.

So the question is a pragmatic one: how useful is the Bayesian approach for understanding cognition, in comparison with (and alongside) other approaches. Many of the chapters in this book outline encouraging developments; and literature in Bayesian cognitive science has been expanding rapidly, suggesting that the approach is often productive. My guess is that this trend is likely to continue—and that attempting to understand cognition without Bayesian ideas in the theoretical toolbox is somewhat perverse. But let’s see what happens over the coming decades.

*S: Well, indeed, the ultimate usefulness of any approach can only be determined by trying it out—and perhaps we can agree that the jury is still out on this one, however promising some of the progress so far may be. But let me finish with a final worry, which makes me doubtful that the Bayesian approach can really be on the right track: and this comes from thinking about neuroscience.*

*Cognition is generated by the brain. Yet this entire book scarcely mentions a single fact of neuroscience. The approach is relentlessly top-down. Try to analyze the nature of the problem the brain is facing (Marr’s computational level) and how it might be solved in principle; perhaps consider algorithms that might approximate the solution (the algorithmic level); and only then worry about how such calculations might be implemented in neural hardware (the implementation level). But, let’s face it, we never got to the implementation level, did we?*

*And isn’t this a crucial step? Aren’t we ultimately reverse engineering the brain, not merely the mind. And surely we can’t reverse engineer the brain without paying a lot of attention to the computational*

*machinery of the brain. Here, neural networks seem a lot more promising—because their basic mode of operation (computing with densely interconnected networks of numerical processing units and learning through adjusting the strengths of connections between those units), seems a lot more promising.*

B: This is more a question of research strategy, rather than substantive debate. Building a link between high-level cognition and low-level neuroscience is inevitably going to be incredibly difficult; and it likely that there will be strong constraints between levels of analysis. So, for example, it has long been pointed out that purely sequential symbolic computations seem incompatible with the slow, parallel hardware of the brain (Feldman & Ballard, 1982).

Trying to build in all constraints at once is terrific, if doable. In some areas, such as aspects of perception and motor control, and perhaps basic process of reinforcement learning (if these can usefully be distinguished from higher level cognition; Chater, 2009; Mitchell et al., 2009) a multilevel analysis seems at least possible. Indeed, this is the project of computational neuroscience, much of which takes an explicitly Bayesian perspective (Doya et al., 2007; Knill & Pouget, 2004).

When we consider areas of higher-level cognition, as is the focus of this book, this goal seems currently to be out of reach. But to get there, we should surely be working piecemeal at each level of analysis, and linking levels together where we can, to try to piece together a complete story. As it happens, progress has been easiest at Marr’s computational and algorithmic levels, so far—connections with neural implementation are not yet well-developed enough to provide neural constraints on reasoning about the physical or social worlds; we don’t yet have a clear sense of how to neurally implement complex symbolic representations; and so on. For now, let’s make progress where we can—and keep channels of communication across levels as open as possible.

To end on an optimistic note, it is worth stressing that the challenge of meeting constraints from multiple levels of analysis, where we can figure out how to do it, is actually likely to speed rather than slow our progress (Christiansen & Chater, 2016b). If we are attempting to solve a puzzle of any kind, more clues are better than fewer—more clues will guide us towards good approaches and eliminate the bad. So when it comes to reverse engineering the brain/mind, we should use constraints wherever we can find them. Bayesian cognitive science, when applied to high level cognition, is not always able to account of constraints from human biology, because it is not clear how neural implementation connects with quite abstract computations problems and algorithms. But the more we can join together insights from different levels of analysis—and from diverse fields, from AI to cognitive psychology to linguistics, philosophy and neuroscience, the more rapidly we will progress. Bayesian cognitive science is, after all *cognitive science*. And it is this interdisciplinary approach to providing an integrated account of the brain as a computational machine that gives us our best hope of understanding the astonishing phenomenon of human intelligence.

# Conclusion

We began this book with a big question: How do our minds get so much from so little? The intervening pages have sketched an answer, using Bayesian inference as a tool for reverse-engineering the mind. Thinking about the ideal solution to the computational problems the mind faces provides a framework for exploring people’s inductive biases, expressed in the form of prior distributions. The challenge then becomes one of finding the right language for describing these prior distributions (drawing on structured representations such as graphical models, logic, grammars, and programs), understanding how they can in turn be learned and deployed when relevant (via hierarchical Bayes, nonparametric Bayes, and metalearning), and how they can be used effectively and efficiently by minds and brains (through sampling and optimization). The chapters you have read are the result of more than two decades pursuing these threads, but they all join back to that first idea of Bayes as a paradigm for inductive inference.

Consistent with the reverse-engineering approach we outlined in our first chapter, the book has worked through multiple levels of analysis from the top down. At the level of computational theory, we offered Bayesian inference as an account of the ideal solution to the problems faced by human minds. At the algorithmic level, we explored Monte Carlo and variational inference as mechanisms for approximating Bayes, and considered how they might be best deployed as a bridge back to the computational level. At the implementational level, we highlighted connections to neural networks and ways in which different algorithms might be expressed in neural circuits. And across all levels we considered what kinds of representations might be needed to effectively and efficiently express the hypothesized operations, with all the adaptiveness and robustness that human intelligence displays. The result is a fully integrated picture of how our minds could work – as machines for approximating Bayesian inferences and decisions over hierarchies of flexibly structured representations, carried out via neurally implementable, resource-rational algorithms.

This picture of human minds is equally useful as a lens for understanding AI systems. While our focus has been on human cognition, our stated goal is for the ideas that we present here to be equally useful to practitioners of AI. By contrast to the approach we have taken in this book, much of contemporary AI has adopted a more bottom-up approach in which modeling begins by specifying a neural network architecture that is differentiable, together with sources of training data and loss functions, and then intelligence is left to emerge via gradient-based optimization of the parameters of the architecture given the data and loss function. This approach has been remarkably successful in machine learning, but it also has exposed increasingly wide gaps between AI systems and human minds. We would suggest that those gaps are in exactly the capacities that are highlighted by studying those human minds from the top down: inductive biases, representational modularity, hierarchy and flexibility, and rationally adaptive ways of approximating rational Bayesian inferences and decisions.

Our formula for reverse-engineering the mind – Bayesian inference over structured representations approximated sensibly via neurally plausible algorithms – also provides a way to close those gaps. Many of today’s AI systems that are built by training large artificial neural networks, loosely inspired by the brain and without explicit reference to either structured representations or probabilistic inference, can be shown to be implicitly doing something like this – forming representations that can be decoded to reveal analogues of posterior distributions over structured representations that have been learned to make sense of their data. In some cases, given enough data and the appropriate architectures, good-enough approximations to structured probabilistic models and inference algorithms might be learned without explicitly building in these components. The more this program is successful, the more we expect that the concepts and tools in this book will turn out to be useful in understanding how those large-scale neural network systems actually work – so they do not just have to be treated as black boxes. And to the extent that this approach to AI continues to fall short of what we expect from intelligence, if AI designers want to create systems that learn and reason more like people – from less data, in a broader range of settings, more robustly and coherently, and with a more robust and coherent sense of their own uncertainty and ability to use their own computational resources rationally – we also expect they will find value in the tools and ideas we present here, as an explicit guide for engineering and not just reverse-engineering intelligence. Intelligent designers of intelligent machines owe it to themselves, to their

creations, and to all of us to be as thoughtful as possible about how they might design into those systems the right machinery for thinking – the right inductive biases, representations and computations.

In the preceding chapters, we have laid out some of our ideas for how to achieve this synthesis of AI and cognitive science. By better understanding human inductive biases we are better able to construct the targets for our AI systems. By exploring methods such as hierarchical Bayes and metalearning we can see how those targets might be reached partly through experience. By discovering how people efficiently approximate complex problems of probabilistic reasoning we obtain clues that can be transferred to machines. And by exploring the potential of probabilistic inference over representations like graphs, grammars, logic, and programs we have the components for describing human inductive biases clearly, in a form that we can use to create AI systems that come closer to instantiating those inductive biases.

In the second part of the book we also laid out some of the most important ongoing and future directions for this research program. Perhaps the most significant theme is this one: re-unifying AI and cognitive science. There are still going to be significant differences between the fields – human intelligence is shaped by a specific set of computational constraints that are the product of our biological and cultural evolution, and not the same as those that shape AI, in a continually changing economic, technological and cultural landscape that AI itself is reshaping. Yet there remains a significant amount these fields can continue to learn from one another. Reverse-engineering can provide insights that can inform engineering, and vice versa, as long as the scientists and engineers involved speak a common language. We hope that this book provides a dictionary and a grammar – or at least a phrase book – for such a language.