CHAPTER 1

Introduction
Toward a Cognitive Science of Belief

Joseph Sommer, Julien Musolino, and Pernille Hemmer

1.1 Introduction

Beliefs play a central role in our lives: They lie at the heart of what makes us human, they shape the organization and functioning of our minds, they define the boundaries of our cultures, and they guide our motivation and behavior. It is no surprise then that belief has been studied in a broad range of disciplines, including anthropology, sociology, political science, economics, philosophy, and psychology. However, trying to determine what beliefs are, or what they ought to be, is far from obvious. For example, the philosophical literature on belief offers a dizzying array of possibilities, including one form or another of representationalism, dispositionalism, interpretationism, functionalism, eliminativism, instrumentalism, atomism, holism, internalism and externalism. The psychology of belief has long history from introspectionist studies (Okabe, 1910) and theorizing (James, 1889), to early AI (artificial intelligence) models (Abelson, 1973; Colby, 1964; 1973), and belief has been studied across numerous psychological domains (see Porot & Mandelbaum, Chapter 3, this volume). Today, it seems that interest is coalescing around belief as a field of its own (Connors & Halligan, 2015). Still, psychology has often struggled to assemble a comprehensive treatment of belief (Egan, 1986). Part of the difficulty may be that the range of phenomena, patterns, and distinctions that seem to fall within the scope of a theory of belief is arresting.

To see this, consider the following observations: (a) beliefs can have different origins. They can be formed through direct sensory experience, through interactions with others, or via written or audiovisual information channels; (b) beliefs can be held at different levels of awareness. While some of our beliefs are accessible to conscious inspection, others may not be; (c) beliefs can be held with different levels of conviction. People can be very confident about some of the things they believe and

much less so about others; (d) beliefs vary in their susceptibility to change. Some beliefs appear to be very stubborn and difficult to change while others seem much easier to modify or even abandon; (e) beliefs vary in their generality and scope. Some beliefs apply to single objects or individuals while others apply to entire classes of entities; (f) beliefs vary in the extent to which they guide our behavior. Someone might be afraid of flying and avoid doing so and yet hold the explicit belief that air travel is perfectly safe; (g) beliefs can produce a range of emotional effects. While some beliefs are benign or even helpful, others can have dire emotional consequences; (h) beliefs vary in the degree to which they are shared by other people. While some beliefs are idiosyncratic, others can be quite widespread and perhaps even universal.

This diversity has led to disparate conceptions across fields and to literatures that often do not make contact with each other, especially in the context of hyper-specialized academic disciplines. Since complex problems can seldom be comprehensively studied within a single discipline, different aspects of the question of beliefs have been studied by different investigators, using different approaches, and yielding different and sometimes seemingly contradictory results. This fractured picture calls for a systematic effort to integrate these disconnected lines of research and start a broader dialogue on the nature, role, and consequences of beliefs. This is the goal that we set out to achieve in the present volume. Because beliefs are anchored in minds, it is fitting that the integration we are calling for here be conducted within cognitive science, the interdisciplinary study of mind.

In this opening chapter, we offer a brief introduction to some of the perennial questions posed by the study of beliefs. In section 1.2, we begin by discussing the nature of beliefs and their place in the study of mind. Section 1.3 asks why biological organisms – human beings at least – should have beliefs in the first place. Section 1.4 briefly discusses research on belief formation and updating which seems indicative of irrationality. These findings, as well as evolutionary considerations, have led to proposals about functional roles for beliefs other than aiming at truth. Central to these issues is the question of whether, or the extent to which, human beings can be regarded as rational. In section 1.5, we review considerations about epistemic trade-offs, computational constraints, and the effects of prior knowledge, which suggest that ascriptions of irrationality may be tempered by the complexity of the problems that belief systems face. We close this section with a note on irrationality. Finally, in section 1.6, we offer a brief overview of the content of the book.

1.2 What Are Beliefs?

The nature of beliefs has been debated for centuries and philosophers have yet to reach a consensus definition (Quilty-Dunn & Mandelbaum, 2018). As mentioned above, some of the difficulty stems from the variety of phenomena that we intuitively label as belief. Beliefs are often unequivocal assertions of flat-out belief (Frankish, 2009), where one either fully believes or disbelieves a proposition (e.g., "The sky is blue"). On the other hand, they may also vary in the degree of confidence we attach to them (e.g., "It may rain tomorrow") (Ramsey, 1926). Additionally, propositions that are actively entertained (e.g., "I need to make a right turn on Pine Street now"), stored in memory, (e.g., "It is safe to walk around my neighborhood at 7:00 p.m."), or generated on the fly (e.g., "This guy just tried to cut me off!") all seem to intuitively qualify as beliefs. Even perceptual processes have been argued to be guided by beliefs, *qua* innate or early developing assumptions about features of the environment that aid in recognizing stimuli or in selecting the most likely percept (Goldman, 1986; Pinker, 1997; Spelke & Kinzler, 2007). To make things worse, people often behave as if they hold beliefs that they never explicitly consider. These implicit beliefs may come in the form of unconscious biases that yield behavior contrary to overtly expressed beliefs. Given this heterogeneous collection of phenomena, it is perhaps not surprising that the ontology of beliefs remains elusive.

For philosophers of a certain persuasion, beliefs do not even exist. Instead, beliefs may be regarded as explanatory fictions that account for observed behavior (Dennett, 2017) or mere linguistic expressions without meaningful referents (Churchland, 1981). Among those who accept the existence of beliefs, opinions regarding their nature also vary. Many argue that beliefs are a kind of "propositional attitude" (Quilty-Dunn & Mandelbaum, 2018) that contains at least two parts: a proposition, *P* (e.g., it is raining), and a mental stance, or attitude that one holds with respect to *P*. When someone holds a belief – for example that it is raining – the attitude held towards *P* is its presumed veracity. This is to be contrasted with other attitudes one might hold toward *P*, such as desiring or hoping that *P*. Yet others suggest that belief must be identified with a type of behavioral disposition. One such view is that beliefs are a disposition to behave as if *P* is true or to assert *P* in the appropriate circumstances (Bain, 1872). For example, the belief that it is raining would simply mean possessing a disposition to assert as much or to grab an umbrella before going outside. More elaborate dispositional accounts identify belief with matching a stereotype for believing the proposition in question

(Schwitzgebel, 2002). On this view, believing that there is beer in the fridge means possessing a set of dispositions that might include looking in the fridge when thirsty, offering beer to visitors, not adding beer to one's shopping list, etc. Possessing all or any one of these dispositions is not necessary, as long as one's dispositions are similar to those stereotypically associated with the belief.

It should be noted that the longstanding difficulty in defining belief should not force us to adopt a particular view, nor should it necessarily be regarded as an obstacle to scientific investigation. As Fodor wrote, "pre-theoretically, we identify mental events by reference to clear cases. Post-theoretically, it is sufficient to identify them as those which fall under psychological laws" (1975, p. 4, fn 2; see also Fodor, 1968, pp. 10–11, 143). In other words, our pre-theoretical notions of beliefs are sufficient to focus our attention on certain questions and guide investigation, even if scientific notions of beliefs end up departing from these pre-theoretical conceptions once explanatory theories become available (see Stich, 1983).[1]

In gathering chapters for this book, we chose an inclusive definition of beliefs so as not to artificially impose a priori limitations on the scope of a more mature scientific theory of beliefs. Chapters in the present volume discuss representations ranging from explicit propositions to tacit Bayesian priors. As such, our broad definition includes any stored or generated information that is used to aid perception, action, or cognition.[2] We should be pleased to discover that this definition is in error, as this will mean that the science of belief has progressed toward sharper empirical distinctions. There are other reasons to adopt a broad view. A mature science of beliefs should not merely delimit the boundaries of the phenomenon; it should also be concerned with antecedents and consequences of beliefs. For example, if implicit representations or tacit priors are not deemed to be proper beliefs after all, they may still be of interest for their capacity to generate beliefs or assist them in guiding action (cf. Stich, 1978).

[1] It is worth noting that our naïve intuitions about a broad range of natural phenomena have provided a very useful scaffolding for the development of scientific theories. However, once scientific concepts mature, they all too often become perplexing to our intuitions, an important point already noted by Hume. Commenting on Newton's ideas and the demise of the intuitive mechanical philosophy of the Cartesians, Hume (1778, p. 542) pointed out that "While Newton seemed to draw off the veil from some of the mysteries of nature, he showed at the same time the imperfections of the mechanical philosophy, so agreeable to the natural vanity and curiosity of men; and thereby restored her ultimate secrets to that obscurity, in which they ever did and ever will remain."

[2] A view of belief that is not explicitly propositional may account for animals' and pre-linguistic children's competencies (Churchland & Churchland, 2013). Note, however, that having language may not be necessary for the ability to express propositions in a "language of thought" (see e.g., Fodor, 1986, p. 19).

1.3 A Cognitive Perspective on Beliefs

Though there is presently little agreement on a definition of beliefs, some aspects of the complex nature of beliefs might be better understood by examining beliefs through the lens of cognitive psychology. A cognitive approach raises the prospect that beliefs might differ due to the interaction of multiple psychological processes with computational principles underlying belief formation and storage. This perspective may help explain some of the complexity underlying beliefs discussed in section 1.2.

Perhaps the obvious place to begin an analysis of the psychology of belief is with the question of what benefits beliefs confer on a cognitive system. The primary advantage of having beliefs seems to be the ability to construct and reason about internal representations of the external world instead of being forced to learn from direct experience (Campbell, 1974; Dennett, 2008). For example, thinking counterfactually (Evans & Over, 2004) and simulating outcomes can help us construct better plans and select adaptive actions. However, these abilities do not seem to account for the quantity of diverse beliefs that people come to hold, many of which are not readily associated with actions (Abelson, 1986).

One reason for possessing many beliefs is suggested by Newell's (1994) preparation–deliberation trade-off: when one is faced with a problem, there are two broad options, either come prepared in advance (because say, the problem was encountered in the past), or calculate a solution. Newell (1994; see pp. 102–107) observed that two systems can achieve equivalent performance if one stores more knowledge and the other spends more time calculating. However, within a single system, deliberation cannot be easily improved because of fixed computational capacities like processing speed, but more knowledge can usually be stored (see also Sperber & Wilson, 1996, p. 47). This means that most improvements within a given system will come from learning and encoding new knowledge and procedures. This is especially true for an organism facing real-time constraints on decision-making and action, as computation is likely to take longer than accessing memory. We might therefore expect people to possess a large body of stored beliefs to reduce computational time and effort.

The preparation–deliberation trade-off would seem to imply that the larger the set of stored beliefs the better. However, there are also reasons to limit storage. An agent with a large set of *explicit* beliefs can face severe problems in updating (Janlert, 1987). Upon encountering new information, the agent will need to iterate through its beliefs, determine which are inferentially impacted by the new evidence, and revise each one. As the

number of explicit beliefs grows, this process becomes computationally expensive and eventually intractable. Say an agent has a representation of several items on a table and explicitly stores beliefs about their directional relationships, e.g., a book is to the left of a glass and to the left of a pen, etc. If the book is moved to the right of the table, each of these explicitly stored relationships will become obsolete and require correction. In contrast, if directional relationships are left implicit, they can be calculated as needed, removing the need to keep the representations current (Bobrow, 1975).

Similarly, if many beliefs are generated from a few core beliefs, altering just one of these will de facto update all the beliefs it composes. For example, if one learns that conservative politicians are opposed to a large government, rather than applying this new belief to every known conservative, only a single belief about conservative positions needs to be updated. This belief can generate the information for application to any specific politician (assuming this knowledge is later activated). This approach also gracefully deals with learning about new conservative politicians. If the belief about opposition to a large government is explicitly applied to each known conservative politician, it may not be applied when a new conservative is encountered. On the other hand, generating this information from a single belief about conservatives can flexibly apply it to new cases (Woods, 1975, p. 73). Implicit beliefs may represent exactly this solution, as they can be inferred from other knowledge and do not need to be stored in explicit form (cf. Gallistel & King, 2009, pp. 58, 208). Explicitly storing only a subset of possible beliefs may make updating more tractable (Sperber & Wilson, 1996, p. 85).

Which beliefs should be stored then? Those that can generate many other beliefs are good candidates for storage. In other words, one might store premises rather than conclusions. Additionally, stored conclusions may cause problems if the premises that generated them are later revised, leaving the false conclusions preserved in memory. For example, say conclusion *q* is initially inferred from premise *p*, but at a later time, *p* changes such that it no longer implies *q*. If *q* was committed to memory initially, it may persist later, even if it does not follow from *p* at that later time. However, there may be occasions where storing a conclusion is worthwhile. If a conclusion is derived from disparate sources of knowledge which might not be assembled in working memory on later occasions, it may be worth committing to memory (see Sperber & Wilson, 1996, pp. 106–107). For example, if premises *p1*, *p2*, *p3*, and *p4* are all required to infer *q* and are not likely to simultaneously come to mind, one might want to store conclusion *q*. Such conclusions may persist in memory after

the premises that generated them have been forgotten, leaving burned-in (Doyle, 1979) beliefs that are held for reasons that have been lost (see also Harman, 1986, pp. 41–42). Additionally, beliefs might be stored in accordance with their use, with a preference for frequently relevant beliefs. This may be achieved for free because commonly referenced beliefs will present themselves to memory more often, increasing their probability of storage (Sperber & Wilson, 1996, p. 77). Belief usage might also be expected to correlate with their generality, as more general beliefs will tend to be relevant across a wider range of situations.

In the case of extremely common situations, beliefs may become overlearned, receding out of conscious awareness, as do most habits. This phenomenon is analogous to the "chunking" of behavior (see Simon, 1996, p. 90) which allows frequent behaviors to be compiled and activated by environmental cues (cf. Schneider, Dumais, & Shiffrin, 1982). For example, routines like tying one's shoes are laborious and explicit when first learned, but become automatic with practice, perhaps in a process akin to running a program through an optimizing compiler (Pylyshyn, 1986; see also Abelson, 1973, p. 310).

Chunked inferences may underlie Frankish's (2004) concept of *implicit belief*. Frankish makes an important distinction between two closely related notions that he calls *tacitly* and *implicitly* accepted beliefs. *Tacit beliefs* are those that can be generated from prior knowledge even if they have never been explicitly considered, e.g., the proposition "1,013 is less than 8,927," which would likely be accepted as a belief by someone presented with it for the first time. In contrast, *implicit beliefs* serve as chunked background inferences. For example, when thinking about going out for dinner at night, a person familiar with their neighborhood can assume that it is safe to do so without consciously considering the issue – a compiled inference. However, while away on vacation or in an unfamiliar location, the safety of the neighborhood may not be taken for granted. In this case, more explicit reflection may be necessary.

The picture that emerges from these considerations is that an agent may be expected to have a set of overlearned and chunked beliefs that remain implicit and guide behavior in extremely common situations. For less recurrent events, explicit beliefs are predicted to be those that cannot be easily generated from prior knowledge, that generate inferences themselves, and that are applicable across situations. For rare problems, tacit beliefs may be freely generated and then forgotten to avoid problems of too many explicit beliefs. If this emerging picture holds, the diverse nature of beliefs might be due to the variety of problems faced by the system along with

principles of knowledge organization such as compiling common inferences into chunks and avoiding unnecessary explicit storage. In a similar vein, which cognitive processes are appealed to, e.g., whether one is referring to memory retrieval or the generation of novel inferences, may yield different beliefs with different qualities, making a single definition difficult to find.

1.4 The Mechanics and Functions of Beliefs

If a cognitive approach sheds light on the complexity of beliefs, the next question that may be asked is how should these beliefs work? A reasonable assumption might be that beliefs are meant to accurately represent the world in order to appropriately guide adaptive behaviors (Fodor, 2000, pp. 66–68). However, much of the psychological literature on beliefs has yielded results that appear to conflict with this assumption. The belief literature has largely been concerned with questions of how people search for and process evidence, whether reasoning is motivated, and whether belief updating is proportional to the strength of the evidence.[3] Many findings in this literature are suggestive of biased search, processing, and assimilation of evidence (for reviews, see Kunda, 1990; Nisbett & Ross, 1980), which raises further questions about whether humans are rational or irrational. In light of the prevalence of apparently irrational behavior, some have proposed theories suggesting that the purpose of beliefs is not to form an accurate model of the world. These include several instrumental or functional notions of belief according to which beliefs serve personal and social goals which may not be truth-oriented. In this section, we briefly discuss the mechanics of belief updating, empirical evidence suggestive of irrationality, as well as the evolutionary case for functional theories of belief.

1.4.1 The Mechanics of Beliefs

The mechanics of belief includes how beliefs are updated, how people search for new evidence, as well as how they reason about evidence that supports or opposes their prior beliefs. We first consider how beliefs are updated. Beliefs should, of course, be sensitive to evidence. The normative model for updating a hypothesis upon receiving new evidence is Bayes' rule. Given some prior hypothesis *H* and new evidence *E*, Bayes' rule

[3] In the remainder of the chapter, we refer to these processes collectively as belief processes.

computes how much more likely H is to be true after learning of E. Say the hypothesis is that it rained the previous night, and the evidence is that the ground is wet. Belief in H will depend on three factors. First is the *prior probability* of H – if the setting is a desert, it is a priori less likely to rain than in a forest. Next is the *likelihood* of the evidence if H is true. Given that it rained, how likely is it that the ground would be wet? This may be quite likely, but not certain, as the rain could have been blocked by a barrier. Finally, evidence is less diagnostic when E is expected regardless of whether H is true. For example, if someone nearby is watering the lawn, the ground could be wet even if it had not rained.

Formally, Bayes' rule calculates the *posterior probability* that a hypothesis is true given the evidence, denoted $p(H|E)$. The prior probability of H is represented as $p(H)$; the likelihood of the data *given* that H is true, as $p(E|H)$; and the probability of E is p(E). Bayes' rule can be given as:

$$p(H \mid E) = \frac{p(E \mid H) * p(H)}{p(E)} \quad (1)$$

Given the axiomatic rationality of Bayesian belief updating (Cox, 1961; de Finetti, 1970/1974), as well as the remarkable match between human behavior and the predictions of Bayesian models across cognitive domains (Anderson, 1990; see Chater et al., 2010 for a review), it might be expected that human belief updating should also be governed by Bayes' rule.

It is important to note that most Bayesians are not committed to the brain implementing Bayes' rule at Marr's algorithmic level (Anderson, 1990; Griffiths, et al., 2012), a level of analysis which attempts to specify the nature of the algorithm employed by the cognitive system (Marr, 1982). Instead, the assumption is that Bayesian explanations operate at Marr's computational theory level, which asks about the goals of the system and attempts to explicate the problem that the system is trying to solve. The algorithms underlying human behavior need not be explicitly Bayesian, but could approximate the normative Bayesian solution (Anderson, 1990; Chater et al., 2020).

In contrast to the theoretical rigor of Bayes, decades of findings suggest that human beings do not update their beliefs rationally. People selectively search for evidence that supports their previously held beliefs (Snyder & Swann, 1978; Wason & Johnson-Laird, 1972). They ignore evidence that should be relevant (Smedslund, 1963), and when presented with information that runs counter to their beliefs, people often engage in motivated reasoning to selectively disconfirm the incoming evidence (Kunda, 1990). Additionally, when people are exposed to mixed evidence supporting two

sides of an issue, they show biased assimilation of evidence, accepting the evidence that supports their views and disregarding opposing information (Lord, Ross, & Lepper, 1979; Taber & Lodge, 2006).

Even in the face of apparently overwhelming evidence, people sometimes continue to believe (Anderson, Lepper, & Ross, 1980; Ross, Lepper, & Hubbard, 1975). In "belief perseverance" studies, participants are often presented with false feedback, suggesting they have performed poorly on a task (e.g., distinguishing between real and fake suicide notes). They are then let in on the initial deception and told that the feedback they received was fabricated. In spite of that, participants often cling to the belief that the feedback was an accurate reflection of their ability.

Findings such as these have led many to conclude that humans are fundamentally irrational (but see Anderson, 1990; Gigerenzer, 2000; Oaksford & Chater, 2007 for different conclusions). Others have even argued that these findings indicate that beliefs are not meant to accurately reflect the world, but are better understood as serving some instrumental, rather than epistemic function.

1.4.2 The Functions of Beliefs

If the purpose of beliefs is not to form an accurate model of the world, then to echo Smith, Bruner, and White (1956, p. 1), "Of what use to a man [*sic*] are his opinions?" It is beyond the scope of this chapter to fully survey the approaches dedicated to answering this question, but we present a brief overview below.

From an evolutionary perspective, beliefs matter indirectly and only insofar as they are tied to actions. Evolution cannot select based on an organism's internal states, but only on their resultant behaviors. If a true belief does not result in adaptive actions, it offers no advantage to its holder. Likewise, if an irrational belief reliably leads to actions that increase evolutionary fitness, it may be selected for, regardless of its truth value (but see Fodor, 2000, pp. 66–68 for an opposing argument).

There are various ways in which the evolutionary benefits of an action may be decoupled from a belief's correspondence to reality. For example, if one has unjustifiably high self-confidence, this may result in increased mating opportunities. If everyone in a community, or a particularly prestigious group member, has arrived at an incorrect belief, it may be more advantageous to conform than to dissent. As Goldman (1986, p. 98) observes, the opposite is also true: there are true beliefs that are not necessarily aligned with survival, such as an accurate understanding of