CHAPTER 3

The Science of Belief: A Progress Report

Nicolas Porot and Eric Mandelbaum

3.1 Introduction

Believing is like breathing: it happens so frequently we scarcely notice its occurrence. We avoid places where we believe ourselves to be unwelcome, we arrive in places we believe our dates to be headed to, and we vote for candidates who we think agree with our values. We can believe that God is female, or ungendered, or a very powerful otter, or does not exist. We can believe in God (i.e., put our faith in God), the goodness of humanity, or the calming eventuality of the heat death of the universe. But despite belief's ubiquity in our everyday view of the mind, one might wonder what work belief really does in our theories of the mind. Is belief merely a state in folk psychology or does it have a place in cognitive science too? Is the notion of belief merely a stopgap to be used until more concrete notions arise from psychology or neurobiology or neurochemistry or physics? Here we review the data for the role belief plays in cognitive science. As in everyday life, though explicit talk of belief in cognitive science as such is scant, its tacit usage is widespread. Below, we describe how widely entrenched belief is in cognitive science. We then offer a snapshot of the current best science of belief. In the end, this chapter aims to establish that there is already a robust science of belief, a thesis which will turn out to be surprising to many, even though the evidence for it has been hiding in plain sight.

We contend that this science is mature enough to supply concrete models of (some of) beliefs' functional role, specifically belief formation, belief storage, and belief updating. Producing defensible empirical models of belief suggests that there really is a place for folk psychology in our science of the mind. The more precise the model, the more it allows

The authors thank Joseph Bendaña, Elizabeth Edenberg, Bennett Foddy, Dan Harris, Marc Porot, Jake Quilty-Dunn, Kate Ritchie, Fiona Schick, and Wayne Wu for helpful discussions.

theorists to argue about what features of the folk kind are core, and which more inessential, in cognitive scientific theorizing.

3.2 Philosophical Models of Belief

Beliefs are a type of propositional attitude (that is, an attitude that expresses a proposition). Traditionally, propositional attitudes have been characterized as mental relations one bears to propositions. Take the belief that birds are smart. This belief is a relation to a proposition, the one expressed by "birds are smart." One can bear many relations to that proposition: one can hope that birds are smart, or doubt that birds are smart, or wish that birds were smart, or imagine that they are, and so on. So, in order to pin down which of these relations beliefs are, we will need a more descriptive characterization.

Historically, philosophy has had two major theories of belief *qua* propositional attitude, both different varieties of functionalism.[1] Common-sense (or "analytic") functionalism and psychofunctionalism. The common-sense functionalist holds that attitudes (and mental states in general) are defined in relation to common-sense platitudes held by the community. For example, if you see an elephant grazing in front of you, we can safely assume that you'll believe that there's an elephant in front of you, that elephants can graze in this area, that elephants exist, that if the elephant fell onto you your day would get a lot crummier, that you'd probably be an inferior at basketball after the accident, that you'd feel even more uncomfortable at the circus in the future, that you'd have a story that would set you apart from most of your friends, that your chances for a lifetime made-for-TV movie have increased considerably, etc.[2] True generalizations about the state don't necessarily matter for the characterization – only the widely known platitudes – the ones that everyone knows

[1] An alternative view, which we discuss only briefly here (see footnotes 8 and 16), is *interpretationalism* (Dennett, 1987, 1991): the view that a complex system has beliefs and other intentional states just in case attributing to them those states usefully captures patterns in, and allows us to make predictions about, otherwise complex behavior. Which beliefs an agent (or system) has thus depends on which beliefs best play this explanatory role. Interpretationalism is inherently agnostic about the psychological states underlying belief, and more generally about belief's role in causing behavior.

[2] The common-sense functionalist would apply the same reasoning to each type of mental state (see, e.g., Lewis, 1980). Take pain. It'll be characterized by platitudes such as if you're in pain, you'll try to get out of it, in general you'll avoid pain, if you drop an anvil on your foot you'll be in pain, if you drop a tuba on your foot you'll be in pain, if you have surgery on your foot without an anaesthetic you'll be in pain, and so on, boringly, for seemingly ever. Of course, all of these generalizations contain a *ceteris paribus* clause (as will every generalization or law discussed in this article, and in the special sciences in general. For more on *ceteris paribus* laws, see Lange, 2002).

(or could easily know) and that everyone knows everyone knows, etc. For the common-sense functionalist, what it is to be a belief is to be used in the generalizations like the elephant ones above.

In contrast, the psychofunctionalist characterizes propositional attitudes like belief by using empirical generalizations discovered by cognitive science. Belief is defined by its inputs (e.g., it's a state often caused by perception (Mandelbaum 2015b, 2019) or inference (Quilty-Dunn & Mandelbaum 2018b)), its outputs (e.g., it causes motor behavior), and its relations to other psychological states (memory, attention, imagination, and the like). These relations – and how they are instantiated (e.g., the particular way in which beliefs are acquired by perception – may turn out to be quite surprising. That's because while folk psychology gives us a rough idea of what mental states there are, this conception is just where inquiry begins. Discovering the true nature of these states can only be accomplished by doing some cognitive science. It may turn out that there are no empirical generalizations to be had about a particular attitude. Say there are no psychological laws about, or involving, *hope*. If not, then for the psychofunctionalist hope is not a projectable predicate in cognitive science.[3]

The psychofunctionalist thus likens the discovery of belief to, say, the discovery of gold. We have some prescientific idea of what gold is, and laypeople pick it out by using its superficial properties. Then, by placing that prescientific idea under scientific scrutiny, we can tell if bits of gold share some underlying nature, and if so, what that nature is. To do this, we need to go further than merely using our basic senses; we need to bring in methods of discovery. In the case of gold, we employ the tools of chemistry; in the case of belief, the tools of cognitive science.

Psychofunctionalism posits that there are law-like regularities that characterize the functional role of belief (Quilty-Dunn & Mandelbaum 2018a). But why think there are such law-like regularities in the first place? After all, there is no general science of propositional attitudes as such (yet?), and there do not appear to be any serious empirical inquiries of other particular attitudes, e.g., *wishing*. So why think there should be a science of *belief*?

As we see it, there are two main reasons. First, there is the power of folk psychological prediction (Quilty-Dunn & Mandelbaum 2018a). People use beliefs to navigate through everyday life: if you believe that giving a talk

[3] In this case hope wouldn't be an entity in cognitive science in the same way that tables aren't ones in quantum physics – there are just no generalizations in either science utilizing either concept. Nothing here strictly entails that cognitive science shows that hopes don't exist anymore than physics shows that tables don't exist.

is necessary for getting a job, and you desire the job, you'll give the talk. Using the concept of belief through applications of the practical syllogism results in accurate predictions of people's behavior. If we use this sort of reasoning to predict your behavior correctly, that gives us some reason to believe in the accuracy of the reasoning. This is just an application of a general principle: when a given posit increases predictive accuracy, we should raise our credence in the existence of the posit. Since using beliefs seem to be a crucial part of how we navigate the world, assuming that beliefs are a real causal factor is just a consequence of our regular methodology.

Second, beliefs figure in explanatory causal structures all over cognitive science, from fields as disparate as cognitive neuroscience, social psychology, cognitive psychology, computational psychology, and neuropsychiatry. These uses motivate a need for an empirically rigorous theory of belief, and offer a starting point for a psychofunctional theory. The widespread occurrence of belief in cognitive science entrenches belief's status as a critical causal variable in the explanation of cognition and behavior. Observing belief *in situ* in cognitive science also tells us a bit about its functional role. Ultimately, we will need a more fleshed out theory, one which includes robust generalizations about the psychology of belief (which are rapidly emerging – we return to this in section 3.4). In the next section, though, we simply stress the ineliminability of belief across various fields of cognitive science, and we give a sense of the kind of role it plays in some of the fields in which it appears.

3.3 Belief in Cognitive Science: An Overview

3.3.1 Placebo Effects and Reward Prediction Error Learning

Pain is the body's response to perceived injury – it alerts the agent as to where, what type, and how much damage has occurred. But pain perception is not merely a bottom-up process: one's beliefs about pain greatly affect its phenomenology. Although general estimates of placebo effectiveness are hard to come by (domain specificity plays a role, as well as individual differences), placebo analgesia is effective in approximately one-third of patients (Beecher, 1955; Ossipov, Dussor, & Porreca, 2010). Complementarily, one can also find nocebo effects, where one's belief that they are about to experience pain causes subjects to experience pain even though they are subjected to non-painful stimuli (Colloca & Benedetti, 2007). Culturally held beliefs about superficial features of objects affect pain modulation in predictable ways. For instance, the color of a placebo pill will change how users feel after ingestion, with red pills acting as stimulants, compared to blue ones (Jacobs & Nordan, 1979). Similarly, price affects clinical outcomes: the more one pays for a placebo, the more effective it is (Tracey, 2010; Waber et al., 2008).[4]

A skeptic may argue that these effects are just response biases: perhaps taking a red placebo just makes people report that they feel higher levels of arousal, while they actually feel the same. The response bias hypothesis naturally leads to the prediction that lower-level machinery should remain unchanged in placebo cases. And placebos do affect low-level mechanisms, such as the circuitry involved in reward prediction error learning (Gu et al., 2015). For example, when smokers (a) smoke cigarettes with nicotine and (b) believe they are smoking cigarettes that contain nicotine, they show elevated activity in the bilateral ventral striatum. That activity is indicative of the value signal at play in reward prediction error learning. Yet smokers show the same activation – and the same learning curve – if they smoke nicotine-free cigarettes that they believe contain nicotine. Placebos did not only cause false positives, but false negatives too: subjects who smoked real cigarettes that they falsely believed to be nicotine-free showed attenuated striatum activity. The reach of belief was widespread, affecting both the value signal and the reward prediction error signal separately (ibid., p. 2541).

Talk of bilateral ventral striatum activity is only of interest here because it is indicative of dopaminergic pathways involved in reward prediction error learning. In other words, your beliefs modulate how dopamine is released in your unconscious learning pathways. If beliefs can have computational consequences that far down, it is hard to say they are merely some useful fiction to be excised by a future cognitive science. On the contrary, beliefs instead look to be a critical nexus around which the mind pivots.

There is nothing special about nicotine. Similar results have been shown for cocaine (Kufahl et al., 2008) and alcohol (Gundersen et al., 2008; van Holst et al., 2014). Shockingly, placebo effects have been shown to be even stronger than opioids. The synthetic opioid analgesic Remifentanil – a very fast, short acting narcotic administered intravenously and used in the operating room for deep sedation – has what one might consider unmediated biological effects on behavior. Yet even these effects are swamped by

[4] This datum is reminiscent of the effort justification in dissonance theory (Quilty-Dunn & Mandelbaum, 2018a). There, effort serves as a proxy for preference: the more effort one puts in, the more one infers they like the activity that demanded the effort; here cost stands as a proxy for effectiveness.

one's beliefs about the drug – believing that one was getting the drug doubles its efficacy, while believing one was not ingesting the drug completely obliterates its effects, even as it is still being administered (Bingel et al., 2011). Remifentanil is no garden variety party drug – it is 200 times more potent than morphine and twice as potent as fentanyl (Wiklund & Rosenbaum, 1997). And, while to our knowledge this is the only published effect (so far) of belief completely inhibiting an analgesic, there is plenty of evidence in the medical literature documenting analgesic placebo effects and pain nocebo effects (Watson et al., 2012).

All this is premised on the idea that researchers take belief to be the mental state doing the heavy lifting in the explanation of placebo effects, rather than some other part of a participant's psychology. This claim should actually be fairly uncontroversial. It stems from a paradigmatic feature of the folk concept of belief: when you tell someone that p, all else equal, they form the belief that p (see, e.g., Mandelbaum 2014).[5] When researchers tell a participant that the treatment she is about to undergo will have a particular outcome, or they otherwise suggest as much (for example, by handing her a cigarette), they expect her to form the relevant belief.[6]

3.3.2 Attribution Theory

There are deep commonalities in explanations of placebo effects and attribution theory. In attribution processes – and in dissonance processes in general – what one believes has downstream effects on a host of other, seemingly unrelated processes. Attribution theory, broadly, is the study of how people attribute mental states to themselves and others. One could pick almost any misattribution of arousal study to show the power of belief, for they share a similar logic (Ross, 1977; Schachter & Singer, 1962; Storms & Nisbett 1970). In a striking example, male college students were asked to take a powdered-milk placebo pill before what they were told was

[5] In that sense, the Spinozan theory (see section 3.4.1) is part of folk psychology. Folk psychology doesn't take a stand on the more abstruse aspects of Spinozan belief acquisition (the dual process theory aspects, or the modal strength of the claims) but does capture enough of the central idea (when people hear that p, they tend to end up believing that p) that some of theory seems to be part of folk psychology. In this way one can see the psychofunctional program being carried out – some aspects of the concept are picked out in folk psychology by its functional role, and then the concept is precisified in cognitive science.

[6] It only strengthens the case here to observe that this platitude about testimony and belief is widely accepted outside of placebo research. It reinforces what is tacitly assumed by the idea of a cognitive science of belief: that the folk psychological kind is at least somewhat continuous with the scientific one appealed to in placebo studies. Similar scientific reliance on the folk concept of belief crops up frequently in psychology, including in the cases we consider in the next three subsections.

a study on its effect on memory (Zanna & Cooper 1974). They were told that it induced stress, that it induced relaxation, or that it had "no side effects." After the filler memory task, participants wrote essays against allowing controversial speakers on campus – a *counterattitudinal* viewpoint on campuses at the time. They were told the essays would be read by the "The Ivy League Administrators Association" who were trying to formulate a policy on allowing controversial speakers, thus ensuring that the students thought that their essays would really matter. Some students were asked to participate in this part of the study voluntarily, a manipulation that made it harder for them to rationalize writing the essay.[7] As one might expect based on normal dissonance effects, more of these participants supported banning controversial campus speakers after writing the essays. But not all of them. Those that had received the "stress-inducing" placebo changed their minds about campus speakers as rarely as those who had been explicitly instructed to write the essay. Having been given another attribution base for their unease – the stress-inducing pill – they no longer felt the need to rationalize their feelings and thus had no need to update their views on campus speakers.

The importance of the last datum is not that placebos can affect one's political views. Instead, the crucial point is that the attribution process – an extremely pervasive phenomenon which occurs anytime we want to understand what we are feeling – is shot through with beliefs and unconscious, instantaneous inferences from those beliefs (Bendaña & Mandelbaum, 2021; Quilty-Dunn & Mandelbaum, 2018b). As long as we have a science of attribution, we will need to quantify over beliefs.

3.3.3 Theory of Mind

You watch someone struggle with the pickle jar, then attribute to her the desire to eat pickles. This is an application of theory of mind (ToM), the ability to attribute intentional states to others. Of course, we can attribute beliefs to others, and so, in that way, beliefs figure in ToM. But we must

[7] Participants' reports of their own felt "tension" (as seen in the "no-side-effect" condition) suggested choice freedom indeed induced dissonance. For the completist, high-choice participants were told "I will leave it entirely up to you to decide if you would like to participate in it, but I would be very grateful if you would . . ." whereas low-choice participants were told "During this wait, I am going to ask you to do a small task for this opinion research experiment." Low-choice/low-freedom participants rarely show dissonance effects (as they feel forced to do what they do and don't have to rationalize their choices further) and aren't discussed further here.

also possess beliefs in order to attribute them.[8] Attributing a belief, desire, or other mental state to some agent is having a belief about the mental states of that agent. And while, strictly speaking, ToM discussions often advert to "attributing" or "thinking about" the states of others, it is not at all clear how one could attribute or think about those states without having some kind of occurrent belief about them.

A common test of ToM is the false belief task. Typical false belief tasks involve a "change-of-location" test. An experimenter places an object in a box or other opaque container, and a second experimenter moves the object to another opaque container, or out of view. What is manipulated is whether the first experimenter can be expected to perceive the location-change (for example, she might leave the room before the location-change, then return afterwards; or instead she might stay in the room, looking at the boxes the whole time). Several paradigms have been used to measure whether children, adults, and non-human animals respond differently in these two conditions, but they share a basic logic: if participants respond differently in the two conditions, it is because they manage to represent the first experimenter's cognitive state. In one condition they represent her as having a false belief, and in another, a true belief.[9]

Considerable debate surrounds whether false-belief tasks track ToM, particularly in infants and non-human animals. One common objection, among many, is that they might be measuring attributions of behavioral or psychological features other than propositional attitudes. Consider variations of the false belief task run on chimpanzees, some of which manipulate whether a higher-ranking conspecific than the participant has visual access to a piece of food in a competitive setting. The chimps go after the food when the conspecific lacks visual access, and refrain when the conspecific has visual access. Many have objected that this task does not utilize chimp ToM per se, but instead relies on other kinds of thought about the conspecific. For example, the chimp might merely expect that

[8] For the interpretationalist, the morals here get a bit tricky as attributing a belief to others suffices for the other to have the belief full-stop. So for interepretationalists, it is possible to have a situation where (say) you attribute a belief to me, but no one has ever attributed any beliefs to you, in which case you would not count as having beliefs even though you attribute them to others.

[9] These paradigms, which have been implemented on children, infants, and non-human animals alike, include: violation-of-expectation tasks (e.g., Onishi & Baillargeon, 2005), anticipatory-looking tasks (e.g., Clements & Perner, 1994; Krupenye et al., 2016), "explicit" false belief tests (e.g., Baron-Cohen, Leslie, & Frith, 1985), and spontaneous-helping tasks (Buttelmann et al., 2017; Buttelmann, Carpenter, & Tomasello, 2009).

the conspecific will go after[10] the nearby food (Povinelli & Vonk, 2003), or it might attribute the perceptual state of seeing the food to the conspecific, and then infer that it will go after the food (Andrews, 2017). Or perhaps it attributes *knowledge*, rather than belief, to the conspecific concerning the food's location (Phillips et al., 2020). But expectations and attributions are species of belief. Even on these alternative interpretations, the best explanation of apes' success on false-belief tasks presupposes they have beliefs about their conspecifics – beliefs about going after food, seeing food, or knowing about food. So even granting this style of objection, explaining false-belief task performance should move us to attribute belief to those that pass it. As it happens, this includes infants and several non-human animal species.

Belief is clearly pivotal to interpreting false-belief task results. But the lessons here are more general. First, belief seems to be so central to ToM that it is reasonable to think any test that is diagnostic of ToM will be diagnostic of the capacity for belief.[11] For example, preverbal children and some non-human animals appear to pass anticipatory-looking and violation-of-expectation tests, which do not rely on verbal responses. Independently of whether such tests really do measure a capacity for ToM, the success of non-linguistic organisms on these tasks dissociates linguistic competence from the capacity for belief.[12]

Belief's role in ToM makes it pivotal to psychological explanation generally. ToM is not some negligible, peripheral ability but instead appears to be a core part of many animals' cognitive and behavioral repertoires. ToM is widely thought to underpin much of social cognition; some have even theorized that lacking it is the main cause of autism (Baron-Cohen, 1997). In sum, ToM is central to (developmental, comparative, and clinical) psychology; thus, so is belief.

[10] The "go after" locution is mentalistic and intentional. Surprisingly, it is also the one Povinelli & Vonk themselves use to describe the contents of the subordinate chimps' attributions (2003, p. 153).

[11] Ditto for mind perception, our ability to think of other people, animals, and objects as things that can feel, think, decide, on the one hand, or hurt, taste, and wonder, on the other (Gray, Gray, & Wegner, 2007). It is plausible that folk concepts of planning, communication, and, especially, thought, are closely related to a folk concept of belief. Mind perception for these capacities, like ToM, relies on a folk concept of belief. And, as was the case for ToM, belief is the only means by which one can attribute such folk states. Mind "perception" just is the formation of beliefs about the presence of other minds.

[12] ToM research also supports the idea that the folk psychological kind of belief does not float free of the scientific kind at work in psychology. The kind at work in explaining how a chimpanzee or an infant passes a false belief appears to be continuous with the folk psychological kind.

3.3.4 *Belief in Comparative Cognition*

It might seem obvious that if human psychology adverts to beliefs, a psychology for at least some other animals will, too. But a long tradition in theorizing about animal cognition rejects the obvious. Descartes is often credited with the idea that non-human animals are "mere" automata, lacking mental substance. As such, animals could not have mental states of any kind, much less beliefs. More recently, some philosophers (Davidson, 1982; Stich, 1979; Bermúdez, 2008[13]) have attempted to tie the capacity for believing to natural language competence. This strategy does not deny non-human animal mentality full-stop, but it keeps the spirit of the Cartesian view alive by denying genuine non-human animal beliefs. And thanks to Morgan's Cannon (Clatterbuck, 2016) and the very long reach of behaviorism in the study of animal learning and behavior, researchers studying animal cognition were long loath to use belief-talk in their work.

In recent decades, the obvious has struck back. It is increasingly common for comparative psychologists, cognitive ethologists, and philosophers alike to describe animals as *inferring* conclusions (e.g., Pepperberg et al., 2019 (an African grey parrot); Jensen et al., 2018 (macaques); Tibbetts, et al., 2019 (paper wasps!)); for an overview, see Porot, 2019), *deciding* between alternatives (Rosati & Santos, 2016 (macaques)), *planning* for the future (Cheke & Clayton, 2010 (scrub jays, great apes, and rats)) *possessing* concepts (Howard et al., 2019 (honeybees)), *attributing* false beliefs to others (Buttelmann et al., 2017; Krupenye et al., 2016 (great apes)), and *engaging* in metacognition (Smith, Couchman, & Beran, 2014 (chimpanzees, macaques, capuchins, pigeons)). All of these actions are intentional through and through; it is plausible that all of them bring belief in their wakes also.

It goes without saying that virtually every claim about an animal or species' possessing an intentional capacity has its detractors. For example, as we pointed out earlier, finding alternative explanations for false-belief

[13] Strictly speaking, Bermúdez does not reject that animals may have beliefs: He thinks that natural language so changes the nature of the beliefs one can hold that non-human animals' beliefs are deeply different from ours in nature. For example, for Bermúdez animals cannot, even in principle, have thoughts about thoughts. This brings out a common problem known as chauvinism: How can we characterize beliefs such that (a) they are properly characterized in humans and (b) they can still apply to other possible intelligences (Block, 1980). Generally the problem is taken to be one for psychofunctionalism, but Bermudez brings out that it is also a problem for common-sense functionalism: if one thinks it is a platitude that one can have beliefs about beliefs (which surely seems platitudinous), then, if Bermudez is right, animals cannot have beliefs.

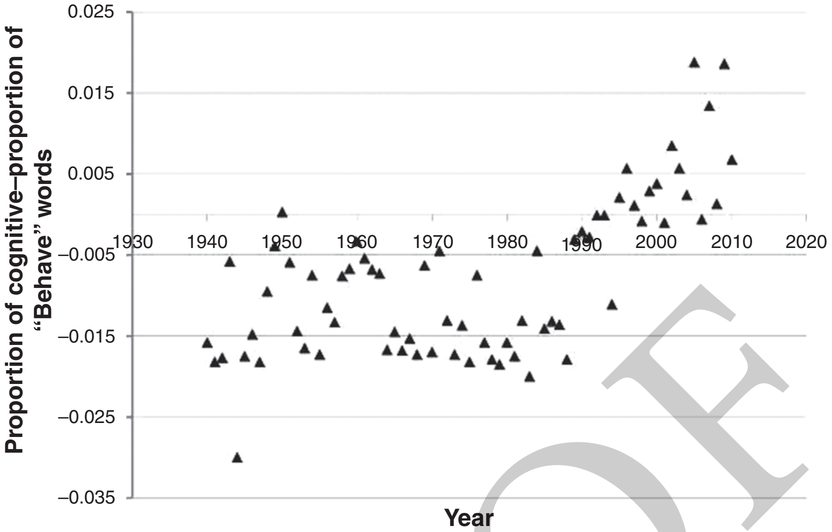

Figure 3.1 Mentalistic titles only surpassed non-mentalistic titles in number in three leading comparative psychology journals starting in the early 2000s.
Reproduced from Whissel, Abramson, and Barber 2013

task success in chimpanzees has become a virtual subfield of its own in comparative cognition (e.g., Andrews, 2012, 2017; Lurz, 2015; Povinelli & Vonk, 2003). Still, this represents a radical shift of the lines of debate. As recently as the 1990s, mentalistic language of any kind in comparative psychology journals was remarkably scarce. Using tables of contents of three journals dating back to 1940, Whissel, Abramson and Barber (2013) found that mentalistic words only *equalled* behavioristic ones for the first time around the turn of the millennium (Figure 3.1). They have never trailed since. The trend toward belief-talk does not reflect a mere change of rhetorical style among some researchers. For great apes generally, and especially chimpanzees, as well some monkey, dolphin, and avian species, non-intentional explanations of many behaviors have become increasingly, implausibly baroque.

The case of belief in non-human animals is an especially striking example of the centrality of belief to cognitive science. Decades of research and centuries of theorizing aimed to avoid talk of belief and related intentional states in non-human animals. Such avoidance was the raison d'être of behaviorist forays into human psychology – it was just patently clear to them that animals could not have these states and, since we are just animals, we should suppose that we do not have them either. Now, we can finally turn this logic on its head: if ever there were a domain of psychology where belief should turn out to be absent, it would be the study of non-human psychologies. That belief-talk has re-emerged in this field for even a few species nicely illustrates how integral belief is to cognitive science.

3.3.5 Belief in Cognitive Science: A Midterm Report

The preceding should give the reader a taste of belief's ubiquity in cognitive science. From comparative psychology, to computational psychology, neuropsychiatry, social psychology, and developmental psychology, belief is ineliminable – not just from our daily folk practices, but from our everyday cognitive science. The four short subsections canvassed a relatively arbitrary bunch – we could have instead focused on entirely different fields, such as delusions, heuristics and biases, terror-management theory, dissonance theory, prospection, metacognitive fluency, stereotyping, self-affirmation, and many others. We suspect this list continues on, for more areas than we are currently aware of.

As belief arises in so many areas of cognitive science, it deserves pride of place alongside such venerable stalwart concepts as memory, attention, perception, and mental representation. However, we have not yet shown the reader much evidence that there exists a science of belief, as opposed to the need to posit belief in scientific theorizing. We will now lay out some empirical results and generalizations of this nascent field, focusing on how beliefs are acquired, stored, and changed.

3.4 The Science of Belief: Belief Acquisition, Storage, Change

3.4.1 Belief Acquisition

Normative epistemological theories – models of how we should believe – generally recommend norms such as: apportion your belief (/credence) to the evidence. If the evidence is equivocal, or non-existent, you should withhold judgment. This intuitive view is enshrined in a picture of belief acquisition called the "Cartesian Model" and it posits that we first entertain a proposition before accepting or rejecting it (Gilbert, 1991; Gilbert, Krull, & Malone 1990). Both acceptance and rejection are seen as active, effortful processes.

But models of belief acquisition need not adhere to all such norms. A competitor view, the "Spinozan Model," posits distinct systems of belief acquisition and rejection. On this view, belief acquisition is effortless, automatic, and mandatory. The processes underlying rejection or endorsement, by contrast, are active and effortful: one has to have the proper available cognitive resources in order to focus on, and reject (or endorse), a proposition. Thus rejection, but not acquisition, can be short-circuited by inducing cognitive load.

One straightforward prediction of the Spinozan Model is that people will reject information less often when put under cognitive load than they would otherwise. Imagine you are presented with a sentence, along with some commentary as to whether it is true or false. Say the sentence is "Corn is the second major export of Angola," and you are told that it is false. If you were presented with that sentence while under cognitive load, you will be apt to misremember it when later queried. You are likely to persist in thinking that it is true that corn is the second main export of Angola. On the other hand, suppose you were presented with the sentence "Coal is the second main export of Guyana," and told that it is true. These same results suggest that if you read this sentence while under cognitive load you would not be apt to misremember the sentence as false – you would persist in remembering it as true. Cognitive load results in misremembering falsehoods as truths, but does not result in misremembering truths as falsehoods. The Spinozan understands this difference as a difference not in encoding, but in rejecting. Both propositions – the one marked true and the one marked false – are encoded regardless of load. That is, both are automatically acquired. When you attempt to add the false tag (Asp & Tranel, 2013), however, you sometimes fail because that takes effort, and so can be short-circuited by load.

The same mechanisms are at work in cases of warnings and retractions. Telling people beforehand that they are about to see falsehoods does not inoculate them from believing the effects of those falsehoods (Wegner, Coulton, & Wenzlaff, 1985; Skurnik et al., 2005). Repeating falsehoods, even ones that the subject knows is false, doesn't stop them from believing them either, regardless of whether the falsehoods are retractions (Seifert 2002; Ecker, Lewandowsky & Tang, 2010; Ecker, Lewandowsky, & Apai, 2011) or concurrent in (as in labeling something fake news). Indeed single prior exposure makes one more likely to believe the headline later on (Allport & Lepkin, 1945; Pennycook, Cannon, & Rand, 2018) even if the headline is preposterous or inconsistent with what else you believe (Fazio, Rand, & Pennycook, 2019; Bendana & Mandelbaum, 2021; Lacassagne, Bena, & Corneille, 2022).

Instead of canvassing the totality of the explanatory work the Spinozan theory can do,[14] we turn to a question that often arises: Why are these states beliefs? Why shouldn't the states so acquired not count as some other, less committal state? One main reason is that the states in question are inferentially promiscuous – they serve as premises in downstream inferences. When someone acquires a proposition they know to be false while under load, they do not just merely parrot it back to an experimenter at some later point. Instead, the information becomes integrated in their overall web of belief and functions just as any other belief, generating actions and increasing their knowledge base.[15]

3.4.2 Belief Storage

How are beliefs stored in the mind? On the realist picture, where beliefs are relations to mental representations, these mental representations have to be stored and accessed.[16] A tempting picture of what such storage looks like proceeds as follows: All the relevant representations reside in a single, holistically connected web in long-term memory. The web is designed to maximize coherence with both incoming evidence and standing beliefs. Since there is only a single memory store, any part of the total information stored can affect any other part.

[14] Examples here include confirmation bias, anchoring and adjustment (one believes the anchor to be correct), source-monitoring errors, self-affirmation, and innuendo, among others. Since the Spinozan theory pays its way by the breadth of effects it can explain, these other data are of some import. For a fuller picture of the scope of the theory see (Mandelbaum, 2014).

[15] For extended discussions on how these states serve as premises in inferences see Mandelbaum, 2016; Quilty-Dunn & Mandelbaum, 2018a; Bendaña & Mandelbaum, 2021. Those papers also contain some other arguments for why these states appear to function as beliefs, for example, they are caused by perception, interact with motivational states to cause action, they are truth apt, and, most importantly, they appear to be the same states at play in other laws about belief.

[16] The reader may wonder: What is the non-realist picture? Popular non-realist views on belief include dispositionalism, with Dennett (1987) and Schwitzgebel (2013) being the two most popular proponents of such views. Dennett's position is hard to pin down, in part, we think, because it is inconsistent. For example, at some points he says beliefs are merely useful fictions, posits that we use to explain behavior but aren't actually causal; in other places (Dennett, 1991, footnote 22, p. 43–44) he claims that beliefs are causal. Schwitzgebel's position is clearer and seems more promising as a research program. He similarly wavers on belief's causal nature but helpfully illuminates the position by comparing beliefs to personality traits. Just as one doesn't quite want to say that being extroverted causes one to make friends easily, so too Schwitzgebel doesn't want to say that beliefs cause behaviors. Instead, Schwitzgebel sees them as pro tem redescriptions of behaviors, mere dispositions to behave. The personality trait talk makes Schwitzgebel seem close to a psychofunctionalist point of view, but he himself is skeptical of there being a robust theory of belief, instead hypothesizing that future neuroscience will illuminate more than any psychology of belief (cf. Quilty-Dunn & Mandelbaum 2018a).

Quine famously defended this "web of belief" picture of belief storage, where belief change transforms the topology of the entire store of beliefs (van Orman Quine & Ullian, 1978). In the periphery of the web are perceptual beliefs, those most easily falsified and least inferentially integrated. Suppose you believe that you are wearing a green shirt. If given evidence that you are wearing a blue shirt (say you realize you are under a colored light), you can change this belief and almost no other beliefs would change with it. Beliefs that are more central are traditionally taken to be ones that are true in the broadest range of possible worlds. So your belief that *modus ponens* is valid, or that 2 + 2 = 4, would be in the center of the web. Falsifying these beliefs would take enormous amounts of evidence, so the story goes, and would completely transform the shape of your web. The web is also thought to be efficient – there is one representation for each token belief, and this representation's location in the web indicates the totality of strength you have in the proposition expressed by the belief.

Human belief storage is not a web. Instead, it looks to be *fragmented* (Bendaña & Mandelbaum, 2021; Elga & Rayo, 2021). Fragmented belief storage contains distinct warehouses of information in memory, called fragments, which are, in the first instance, causally isolated from each other. Different fragments will often carry the same information – different tokens of the same type of belief – but have them associated with differing bodies of information. Fragmentation thus allows for both representational redundancy and a lack of epistemic closure. Representational redundancy is achieved by allowing different tokens of the same type of belief to appear in multiple data structures; lack of closure arises when the premises of an argument are housed in distinct fragments. For example, even if *modus ponens* is "built into the architecture" as a law of thought (so that any time you have the *modus ponens* premises in a perspicuous manner you automatically infer the conclusion of Quilty-Dunn & Mandelbaum, 2018b), lack of closure can be achieved if you believe P in one fragment and if P then Q in a separate one. The lack of interaction between the two premises would block you from inferring Q.

Fragmentation solves a central puzzle in psychological theorizing: How is it that people seem to abhor inconsistencies and yet harbor so many inconsistent beliefs? It allows for a belief system riddled with inconsistencies, as long as those inconsistencies are sequestered from one another in separate fragments. Each fragment is internally consistent and contains no redundant representations. But across fragments, consistency (and simplicity) is not maintained, and representational redundancy and inconsistency are expected. Thus, fragmentation can explain inconsistent beliefs while allowing our belief system to scaffold rational behavior. As long as only one fragment controls our behavior at any time, and as long as our fragments are more or less consistent, rational action will be the norm.

Fragmentation, in the form of functionally distinct memory stores, is already widely accepted in the study of memory: working memory, semantic memory, and episodic memory, as well as other stores more closely linked to perceptual systems. The fragmentation of belief expands this tradition and in doing so can make sense of a wide swath of disparate data. To give the reader a feel for it, we discuss just a few examples of representational redundancy and inconsistent belief.

Take the "wisdom of the crowds within" effect, where people are apt to tap into different bodies of information on the same topic with a mere change in temporal context (Vul & Pashler, 2008). If asked the same question several times (e.g., "What percent of the world's roads are in India?"), participants' mean responses tend to be closer to the truth than any of the individual responses, and the longer you wait between answers the better the averages get.[17] The wisdom of the crowd within effect hinges on being able to tap into independent error. A web of belief view has all of an individual's beliefs connected, so it is hard to say how independent error could be achieved. Fragmentation achieves independent error as a matter of course. Fragments often store different bodies of information from one another. Statistical sampling from one individual at different times will function exactly like sampling across individuals at a time, modulo only differences between the distributions of stored information within and across individuals. Thus, the independent error needed to explain the wisdom of the crowds effect can be achieved within a single individual.

Fragmentation is designed to handle cases where changes in context cause different behaviors without any real informational change. And such cases can be found in reinforcement learning experiments. For example, in *renewal* cases previously "extinguished" associations can re-arise without any new learning, spontaneously reappearing in the presence of their unconditioned stimulus after a change in spatial context. Suppose that a rat put in a circular cage learns to associate a light with a subsequent shock. Suppose then the association is extinguished in a second, rectangular cage (because, for example, the conditioned stimulus – the light – appears

[17] Waiting longer raises the probability of accessing a more completely independent body of information. In the interim, subjects are not getting any new information on the topic as can be seen by the accuracy of their guesses: first guesses tend to be better than subsequent guesses regardless of the time between each guess. And each response is based on the participant's best guess at the time, rather than, for example, on their consulting Google or an almanac.

without the unconditioned stimulus– the shock). A rat that is then put back in the first cage will sometimes show the "extinguished" behavior (Bouton & Bolles, 1979; Bouton & King, 1983; Bouton & Ricker, 1994). The fragmentationalist can explain these cases as ones where we have distinct bodies of information on the same topic, and which fragment is activated changes with the change in spatial context of the animal, as fragments are canonically individuated by context.

A spatial shift isn't necessary to show reinstatement effects. In "spontaneous recovery" (Brooks & Bouton, 1993), a mere change in temporal context – that is, letting time lapse – causes recovery of a previously putatively extinguished association (Bouton, 2006; Mandelbaum, 2015a). Even if one thinks that extinction works by deleting associations from the mind – a difficult position to hold (Bouton, 2004) – the fragmentationalist would expect spontaneous recovery as changes in temporal contexts change which fragment is active (and the information stored in the original association can be found in many different fragments). Similar effects on context for memory storage and recovery have been found in independent avenues of inquiry in computational neuroscience (particularly in the work of Ken Norma's laboratory, see, e.g., Baldassano et al., 2017; Dubrow et al., 2017).

Fragmentation can cover a wide range of other effects (see Bendaña & Mandelbaum, 2021 for a fuller accounting), but, to end, we just want to stress how it is exactly the type of theory that would be needed for a Spinozan mind (Egan, 2008). If every proposition we encounter is initially believed, we will end up believing many contradictions. Producing rational action from a Spinozan mind demands fragmentation to account for our inconsistencies. This interlocking structure, of fragmented storage and Spinozan fixation, is exactly what we should want from a science of belief (and a psychofunctional theory of belief). The laws of belief should be, ideally, more than merely compatible with one another; they should be the natural complements of one another in a well-functioning cognitive system.

3.4.3 Belief Change

There are two robust and distinct research programs in belief change: one stemming from Bayesianism, and another from dissonance theory. Bayesian models of belief are, in general, approximations of near optimal belief updating, changing beliefs in accordance with changes in evidence. The breadth and scope of Bayesianism, at least as a computational methodology, is unparalleled. Where Bayesianism works best is with beliefs that are held relatively dispassionately by the believer. But for updating the beliefs we most cherish, the story gets more complicated.

For beliefs with which we self-identify, rational updating – for example, apportioning and weighing evidence – is not the prevailing norm. People accept and reject information not to maintain epistemic coherence as much as to buttress their sense of self. Belief updating for attitudes of self-importance is largely dictated by a psychological immune system, where counterattitudinal information is seen not just as any new evidence, but instead as a deep psychological threat (Gilbert, et al., 1998; Mandelbaum, 2019). The psychological immune system functions, first and foremost, to help us maintain our most deeply held self-image. Just as our regular immune system does not react to any old bodily injury, but only more serious threats, so too the psychological immune system is not designed to ward off merely feeling bad, but instead kicks in when we encounter serious threats to our self-image. The psychological immune system functions to protect our most core beliefs, the ones that make up our sense of who we are (such as the beliefs that one is a good person, a smart person, and a reasonable person). Believing conclusions that challenge our core beliefs puts us in a state of psychological distress. The psychological immune system remedies this by post hoc rationalizing those conclusions away.[18] The existence of the psychological immune system does not exclude Bayesian belief change for beliefs that we do not self-identify with; rather, the two are distinct pieces of the psychology of belief.[19]

Examples of the immune system are striking. Imagine that you are a John F. Kennedy-assassination conspiracy theorist. If you are shown evidence both for and against the conspiracy theory, you will not do the normatively rational thing and end up tempering your beliefs in the light of the fact that the evidence is equivocal; instead, you are likely to increase your antecedent belief in the conspiracy (McHoskey, 1995).[20] That is, for

[18] Importantly, this is not merely a matter of "hot cognition." The mechanics of the updating are consistent with the computational theory of mind. In general, most rationalizing processes (such as dissonance reduction) are rationally reconstructable from the standpoint of the computational theory of mind, even if the processes themselves aren't particularly rational.

[19] Hence the use of "prevailing norm" in the first sentence of this paragraph. Our claim is not that Bayesian updating isn't also occurring, it's that even if it is, it loses out to self-image based updating for beliefs that are of deep self-importance.

[20] That said, of course, some theorists (e.g., Singer et al., 2019; Nielsen & Stewart, 2020) have identified situations where a good Bayesian would polarize, though we in no way think these situations can whitewash the immune system style polarization and make it rational (though it may be rational in the broader and somewhat pinched sense of rational where avoiding pain is rational).

people who care deeply about the JFK assassination, receiving counter-attitudinal information perversely causes people to increase their belief.

The workings of the immune system have nothing to do at all with conspiracy theories per se. It is just as apt to work off of anything that is central to your self-conception. True believers of any stripe are apt to discount disconfirming evidence because of the crushing effects evidence would have on their way of life, and this holds whether we are looking at anti-vaccers, or just anyone who has received a troubling medical diagnosis.

The immune system also works proactively. When it comes to information gathering, people tend to engage in selective exposure of information. They seek out information that is concordant with their beliefs, and avoid information that is discordant with them (Brock & Balloun, 1967). They look for evidence that they are healthy, and that the mole on their back has not changed shape, or if it has, conclude that that is normal for moles, right?![21]

3.4.4 Beliefs and the Structure of Thought: Bayesianism and Implicit Attitudes

Traditional philosophical analyses have interpreted beliefs as relations to mental representations with truth-apt contents, i.e., relations to sentences in languages of thought (Fodor 1978, 1987). Psychofunctionalism would be well complemented if it could find empirical evidence that cohered with philosophical analyses showing that beliefs have language-like structure; that is, that we have a Language of Thought (henceforth LoT). Somewhat surprisingly, the Bayesian revolution has caused a resurgence in LoT models. The Bayesian "Probabilistic Language of Thought" (PLoT) interprets belief updating (/concept learning) as hypotheses in a language of thought which get updated in line with the canons

[21] There is another sense in which the immune system works proactively. Inoculation theory (Banas & Rains, 2010) – sometimes seen as another branch of the psychological immune system (McGuire, 1964) – sees perseverance as a function of inoculation. For example, giving a subject a weak counterattitudinal argument is seen as similar to vaccinating someone: you give them the weak version so their immune system can realize the threat and build up defenses prophylactically before encountering more serious attacks (Compton, 2020). There is a bit of irony here in terms of the normative appropriateness of the immune system. In Gilbert et al.'s hands, the psychological immune system seems to be an overall practical good, helping us overcome deep traumas and losses in our life (Gilbert et al. 1998). As such, it's an emotional regulatory mechanism. In Mandelbaum's (2019) hands, the immune system is seen as practically positive (as it helps avoid defeats that would bring on existential dread) but normatively inappropriate (as people are increasing credences when they should be decreasing). Inoculation theorists lie somewhere between the two. Inoculation theory is presented as a route to gird oneself against incoming counterattitudinal evidence. But whether such an end is itself positive is at the very least context sensitive: one theorist's normatively suspect belief perseverance is another's properly inoculated response.

of probability theory (Goodman, Tenenbaum, & Gerstenberg, 2014). The PLoT weds the explanatory goods from a recursive LoT (systematicity, compositionality, productivity, deductive inference) to the explanatory goods received from probabilistic reasoning engines (a flexible faculty for generating inductive inferences from sparse and noisy data).

The argument from Bayesianism to the LoT is somewhat indirect. On the one hand, when testing cognitive architectures in concept learning experiments, models that have the full complement of Boolean connectives fare the best at matching human performance (when compared to models that either have no connectives, like exemplar or prototype models, models that are predicated in behaviorist response biases, or even just maximally simple LoT models with only one basic logical connective) (Piantadosi, Tenebaum, & Goodman, 2016). Computational modeling data thus suggests that humans learn concepts based on a LoT-based cognitive architecture with the full suite of first order logic connectives available to it.

Still, there is more to cognition than just learning. Consequently, one may be skeptical of concept learning data as an induction base for how the mind works in general (since, for example, there is no motoric element). However, Bayesianism has been successful in a truly impressive range of tasks.[22] Insofar as most of these models explicitly run on the PLoT, their success counts as evidence in favor of the existence of LoT.

For more evidence that our beliefs are relations to mental representations in a LoT, we turn from computational cognitive psychology to a science often less concerned with mechanism and the structure of thought: social psychology. Consider implicit attitudes. An appealing way of characterizing the "attitudes" under consideration there is to say they are simply beliefs.[23] If that is right, then a few important consequences follow.

[22] For example, Bayesian models have succeeded in learning: sentence to logical form mappings (Zettlemoyer & Collins, 2005), kinship relations (Kemp Goodman, & Tenebaum, 2008); semantic structures (Katz et al., 2008; Mollica & Piantadosi, 2015); general Boolean concept learning (Goodman et al., 2008); folk psychological theories (Ullman et al., 2010); programs such as text editing (Liang, Jordan, & Klein, 2010); causal relations (Goodman, Ullman, & Tenebaum, 2011); integers (Piantadosi, Tenebaum, & Goodman, 2012), analogical reasoning (Cheyette and Piantadosi, 2017); moral rule learning (Nichols, 2021); and modeling motor processes (Rescorla, 2016). Even if Bayesianism didn't explain the workings of most of the mind, as long as some parts of central cognition are Bayesian, the inference from Bayesianism to the LoT would be on solid ground.

[23] Often enough, this is not explicit in the work of social psychologists (by contrast cognitive psychologists and neuroscientists use the word "belief" more readily). "Attitude," however, can be used to mean different things: preferences toward a content, a general valence, mere knowledge of a cultural stereotype (lacking any endorsement of the stereotype), or a belief. In the case of implicit attitudes, it has been suggested that implicit attitude tests measure knowledge of stereotypes, rather

First beliefs are truth apt; associations, on the other hand, are not: one's belief that there is salt and pepper on the table is either true or false; one's association of salt *with* pepper is not the sort of thing that can be true or false. So, if implicit attitudes are beliefs, then implicit attitudes are relations to mental representations that express propositions. Mental representations that express propositions are structured; that is, they are not associative, just as the LoT hypothesis supposes.

The idea that implicit attitudes have propositional structure is tendentious (cf. Brownstein, 2015, 2018).[24] Mandelbaum (2016) and De Houwer, Van Dessel, & Moran (2020) have been the view's strongest proponents (though see the references in footnote 24, as a broader propositional consensus is emerging). Some of the evidence for the propositional view is that manipulations of implicit attitudes proceed via logical and evidential manipulations that sometimes exploit the syntactic structure and logical machinery of thought (e.g., the cases in Mandelbaum, 2016), and, at other times, vary the strength of evidence you receive.[25] For an example of the latter, single pieces of highly diagnostic information (like learning that someone you thought highly of is a criminal) can reverse your implicit attitudes formed over many presentations (Mann & Ferguson, 2015; Cone, Flaharty, & Ferguson, 2021).[26] Thus, the implicit attitudes

than belief in them. We note that in both cases, what is measured is a belief – either a belief that p, or a belief that the culture at large takes it that p.

[24] We, unsurprisingly, do not think it is very tendentious at all (if those who work on attitudes want to excise structured representations from the mind, we wonder how they will reconstruct psycholinguistics). If any interventions in implicit attitudes work due to logical form, that is evidence that implicit attitudes have logical structure. Previously, the arguments from implicit attitudes to beliefs (e.g., Mandelbaum, 2016) went from the observation that attitudes are modulable due to their logical form to inferring that they are structured and then moving from the observation that they are structured to the identification of them with beliefs. Here we attempted to invert that logic and show that regardless of how the attitudes are modulated, they seem to function as beliefs (and then infer from them being beliefs to them having structure). For other evidence of propositional structure of implicit attitudes see Mann & Ferguson, 2015, 2017; Van Dessel et al., 2016; Cone, Mann, & Ferguson, 2017; Kurdi & Banaji, 2017; Mann, Kurdi, & Banaji, 2019; Van Dessel, Ye, & De Houwer, 2019; Kurdi & Banaji, 2019; Hughes et al., 2019; Kurdi & Dunham, 2020).

[25] Other evidence offered for the propositional position is that propositional induction – telling people a piece of information – is a much stronger method of attitude acquisition than repeated associative exposure (see, for example, Kurdi & Banaji, 2019). Similarly, if participants are told that they'll see facts of the form "If you see a green circle, then X is a positive adjective (/negative adjective)," participants only form the relevant attitude if they then see the green circle, thus completing the *modus ponens* inference. Associative theory would predict that mere contiguity would help form the attitude but that turns out to be a misprediction (Kurdi & Dunham, ms).

[26] Admittedly, effects of these manipulations have not been shown to persist in the long term (Lai et al., 2016) but some have at least shown a durability of forty-eight hours (Mann & Ferguson, 2017). Fragmentation can help explain why these effects don't seem to persist (in a similar way to how it explains how spontaneous recovery works; see Bendana & Mandelbaum 2021 for discussion).

literature lends evidence to the venerable philosophical conception of beliefs as propositionally structured strings of mental representations.

We've made a case for robust generalizations about the way that beliefs are acquired, stored, and changed. As mentioned in the introduction, such generalizations are congenial to a psychofunctionalist treatment of belief: they permit a level of specificity that such views have traditionally lacked. At the same time, these are not all the generalizations you might want over belief, and so there are more ways to build out the psychofunctional account. Others on the horizon include: attitude clarity (how you know you hold a particular attitude) (Petrocelli, Tormala, & Rucker, 2007); attitude certainty (how you become certain that you hold the attitude you do) (ibid.); attitude activation (what makes it more likely that a given attitude will come to mind); and need for closure (what makes you need questions to be closed) (Webster & Kruglanski, 1994). For a discussion of these issues, see Mandelbaum (forthcoming).[27]

3.5 Future Directions for the Empirical Study of Belief

It is an exciting time for the cognitive science of belief. The study of belief is well-poised to fulfill the promise of cognitive science by having its many disparate areas of research come together and constrain each other's theorizing. As such, it is an antidote to skeptics about the fruitfulness of cognitive science (Núñez et al., 2019). Bedrock philosophical questions

[27] One might be inclined to thus conclude that humans are accordingly gullible, inconsistent, and rationalizing creatures. But if that is right, then we are incapable of *epistemic vigilance* (Mercier, 2020; Trouche et al., 2016; Sperber, Clément, & Heintz, 2010). Proponents of epistemic vigilance maintain that human reasoning emerged evolutionarily along with communication, and that one of its primary functions is to detect untrustworthy communicators: cheaters and frauds. In support of it, its proponents contend that we are rarely duped by fake news, propaganda, or advertising. They stress that we do not trust just anyone; instead, we carefully track the reliability of sources of information, updating the degree of credence we afford their testimony accordingly.

The scope of human gullibility is a pressing question for both the psychology of belief and the study of communication. But nothing about epistemic vigilance is per se at odds with the mechanisms of belief we have described in this section. The low rates of gullibility you would expect from an epistemically vigilant human population are consistent with the Spinozan account of belief formation. So long as people usually manage to reach the "rejection" phase of belief formation during deliberation (which they are much more apt to do if they think their beliefs are under attack, or have previously been inoculated), it is possible that we are also, generally, epistemically vigilant.

Moreover, even proponents of epistemic vigilance admit that humans are, at least some of the time, led to believe improbable claims through fake news, advertising, and the like. To accommodate such cases, Hugo Mercier (2020) appeals to a distinction between *intuitive* and *reflective* belief (Sperber, 1997). Crucially, the case for this distinction relies much on some of the very same evidence we have levied for belief fragmentation, namely: the widely observed context-sensitivity of belief.

about the nature of belief have become examinable through a wide range of empirical methods. To conclude, we briefly describe avenues for future research: as-of-yet unaddressed questions and promising recent work setting the stage for future research programs.

Earlier, we likened the ontological status of belief to other psychological entities such as memory, attention, and perception. But of course there is not one single mental faculty corresponding to any of these processes: there is working memory, long-term memory, iconic memory, echoic memory; there is endogenous and exogenous attention; there is all manner of different types of perception and attitudes (propositional and otherwise). Perhaps belief can be expected to fractionate into more specific kinds too.

Neil van Leeuwen has offered one means of expanding the ontology of belief-like attitudes (2014, 2017).[28] Van Leeuwen argues many religious and supernatural "beliefs" are not factual beliefs, like the belief that Long Island Sound is polluted, but a kind of "secondary attitude," similar to imaginings, hypotheses, or assumptions for the sake of argument. Like other secondary attitudes, he claims, religious attitudes differ from bona fide beliefs as the former are highly context-sensitive. For example, members of the Vezo tribe in Madagascar speak about ancestors as if they had psychologically survived bodily decay when primed with a ritual context (burying a body in an ancestral tomb) but not when primed with a medical one (dying by malaria in a hospital) (Astuti & Harris, 2008). Van Leeuwen claims that religious attitudes and other secondary attitudes do not have broad cognitive governance over other, non-religious attitudes. For example, Maya-speaking Itza claim humans sometimes transform into animals, yet they also do not worry that meat-eating is potentially cannibalistic (Atran, 2002, pp. 84–86). Finally, religious attitudes are not evidentially vulnerable: learning that a proposed doomsday has come and gone does not typically lead millennial cultists to leave the cult. If religious belief is not factual belief, then belief may need to be separated out into kinds, much like attention, memory, or perception.[29]

[28] Though not the first one: some have argued, for example, that many of the states that look to be beliefs are in fact a different ontological kind: "aliefs" (Gendler, 2008). For Gendler, aliefs are associatively structured, and like irrational beliefs in that they will not update in normatively respectable ways. However, this view faces some serious difficulties (Mandelbaum, 2013). Relatedly, a distinction between tacit and standing beliefs has been adumbrated in many places. One problem (of many) for dispositionalist accounts of belief is the difficulty in distinguishing these two different kinds of belief (the psychofunctionalist posits that one only stores representations for standing, not tacit, beliefs) (Quilty-Dunn & Mandelbaum, 2018a).

[29] A problem for Van Leeuwen-style accounts is that factual beliefs also seem to have problems updating (see, e.g., Anderson, 1983; Mandelbaum, 2019). The question of whether man-made global warming exists is a factual question, and one that leads to a factual belief, yet the updating of it is still stubbornly recalcitrant in the face of evidence. Part of the problem for Van Leeuwen is that

Another outstanding question is whether features – as understood on prototype theory – are best understood as beliefs about categories. A vast psychological literature, calling on an array of experimental paradigms, has shown that categorization is easier for some members of a given category than for others (see, e.g., Hampton, 1995). If asked to name a pet, for example, dogs and cats are more likely to come to mind than ferrets, boas, or piranhas. The most common interpretations of these effects posit stored representations of features. For example, your prototype of birds will have features such as +flies, +feathers, +small, and +cute. What is the status of these features? They do not seem to be mere associates (they are certainly not classically reinforced associates). But they also aren't generally interpreted as having internal structure. Nevertheless, it is unclear if that is a theoretical position taken or a mere typographical convenience. Accordingly, it is an open question whether features function as beliefs about the things they are features of, such that having a prototype bird with a feature +flies is equivalent to believing that birds fly.

Other programs of research include metacognitive effects on belief. Our introspective access to our beliefs is poor, so we are forced to rely on metacognitive cues, such as fluency (Alter & Oppenheimer, 2009, Vogel et al. 2020). One question ripe for exploration is the extent to which fluent access serves as a cue for what you believe (Unkelbach & Rom, 2017; Brashier & Marsh, 2020; Vogel & Rita, 2021). Other questions involve the extent to which people think they have introspective access to their beliefs, and what the exact boundaries of introspective access are.

A current hotbed of inquiry is the relationship between full and partial belief. Much recent formal work has focused on partial belief – credences – and their role in formal epistemological theories and Bayesian cognitive science. But to make Bayesian updating computationally tractable, there must either be some heuristic we use – such as sampling (Vul et al., 2014; Icard, 2016) – to approximate Bayes. What these heuristics are and how broadly they can be used is an in-demand topic. Similarly, the relationship between these states and traditional notions of strength of beliefs is a fruitful topic for future research.[30]

almost all beliefs exhibit some type of context-sensitivity. If belief storage is fragmented, then context-sensitivity/lack of broad cognitive governance is the rule, not an exception. Additionally, there is a robust literature on what happens to millennial cults post prophecy that lead them to increase belief in the cult (so it is not as if the belief does not have wide-ranging consequences. See Festinger, Riecken, & Schacter, 1956).

[30] A related question is whether we need both full beliefs at all, or whether credences can do the work of full belief (for a recent discussion arguing that we need both concepts, full and partial belief, see Weisberg, 2020).

Then there are related questions about individual differences in believers: what makes you more susceptible to believing pseudo-profound bullshit (Pennycook et al., 2015; Tracy et al., under review); what makes you more apt to reject any claims and "nay-say" (Knowles & Condon, 1999); why some people are better than others at forming discerning beliefs (Pennycook & Rand, 2019); why some people have greater preferences for consistency than others do (Cialdini, Trost, & Newsom, 1995), and so on. For a discussion of these issues see Mandelbaum (forthcoming).

Of course, questions of belief acquisition, change, and storage are still open ones of active research – we are still just making in-roads into creating justifiable and testable models. That said, we are already beginning to see crossover from the cognitive science of belief to the real-world issues of fake news, brainwashing, and propaganda (Mandelbaum, forthcoming; Mandelbaum & Quilty-Dunn, 2015; Pennycook, Cannon, & Rand, 2018; Pennycook & Rand, 2019). Once we have begun to understand and operationalize belief, we can start using this notion to better understand how advertisers, politicians, social media, and news feeds prey on our evolutionarily ancient methods of forming and changing beliefs. Some of this work has already begun. Gordon Pennycook, Dave Rand, and colleagues have been examining how social media, fake news, and belief acquisition interact. Their work looks at variables such as how warning labels affect uptake of fake headlines, and how multiple presentations of such headlines affect overall credence (Fazio, Rand, & Pennycook, 2019; Pennycook & Rand 2021). So far, their work has supported Spinozan models (under the guise of "illusory truth" effects – see p. 000).

There are many directions in which to take this research. Cognitive scientists can look at how interventions last over time, and assess the strength of the "sleeper effect," by which we come to believe claims over time that we have already rejected as false (Kumkale & Albarracín, 2004). They can also look at how immune system-based polarization works in social media.[31] Similarly of interest are factors contributing to strength of inoculation. Questions here include: inoculation's time-course (e.g., at what point, post-inoculation, we are least responsive to counterattitudinal arguments); which kinds of arguments induce the strongest inoculation; and whether we see similar effects when we come up with arguments

[31] Work here has already begun: Bail et al., (2018) paid Republican Twitter users $11 to follow a bot retweeting left-leaning content for a month. Their views shifted rightward (rather than to the middle), even by comparison with those of a control group of Republicans that did not follow the bot.

ourselves, rather than being given them. Finally, the "illusory truth" effect is an especially promising avenue for understanding fake news. Seeing a sentence twice makes us more likely to judge it true (Hasher, Goldstein, & Toppino, 1977). This occurs even with sentences clearly labeled "False" (Fazio et al., 2015) and sentences that are very obviously false (Fazio, Rand, & Pennycook, 2019; Pennycook et al., 2018). Illusory truth research has already been extended to fake news headlines (Pennycook et al., 2018). The effect occurred even when participants rated the fake articles' believability as low, and it was not erased by labels indicating articles had been shown false by independent fact checkers. It remains to be seen how much social media misinformation affects the outcomes of large-scale events, such as elections, or what the best way to counteract misinformation is (Pennycook et al., 2021).

Studies like the ones just described touch on a host of new research questions for the study of belief and society. Perhaps the most pressing question at the intersection of the cognitive science of belief and cultural studies is: How much of the way we live our everyday life induces cognitive load? If cognitive load shuts down our rejection processes, then it is of the utmost importance to know when we are under load. If it is just when we are doing cognitive activities as intense as counting backwards from 100 in intervals of 7, or writing an essay, then load will be a relatively rare occurrence. If, on the other hand, load is conveyed through much lighter manipulations – say, scrolling through your phone or merely watching cable news – then our understanding of our modern informational environment will differ greatly (Mandelbaum & Quilty-Dunn, 2015; Hawthorne, Rothschild, & Spectre, 2016). If load can be induced through light manipulations, then propaganda will be far more efficacious than we previously thought.[32] The new science of belief thus holds promise to illuminate questions not just about the architecture of the mind, but of how best to construct our modern informational world.

3.6 Conclusion

The philosophical community has generally treated belief as if it were a mere *façon de parler* (Dennett, 1991). At best, it has been thought of as a prediction mechanism that somehow generates correct predictions; at worst, it's been thought to be a vacuous reification. Here, instead of

[32] In addition, it would provide evidence that our belief updating systems are Spinozan, but not epistemically vigilant (see footnote 27).

arguing about whether a science of belief is possible, we have detailed what the current state of the art is, providing a progress report on a field that is flourishing even if the practitioners of it do not always even know that it exists. As such the study of belief is one of the burgeoning fields of a maturing cognitive science.

REFERENCES

Allport, F. H. & Lepkin, M. (1945) Wartime rumors of waste and special privilege: why some people believe them. *Journal of Abnormal and Social Psychology*, *40*, 3–36.

Alter, A. L. & Oppenheimer, D. M. (2009) Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Anderson, C. A. (1983) Abstract and concrete data in the perseverance of social theories: when weak data lead to unshakeable beliefs. *Journal of Experimental Social Psychology*, *19*(2), 93–108. https://doi.org/10.1016/0022-1031(83)90031-8

Andrews, K. (2012) *Do apes read minds? Toward a new folk psychology*. MIT Press.

(2017) Do chimpanzees reason about belief? In K. Andrews, & J. Beck (Eds.). *The Routledge handbook of philosophy of animal minds* (pp. 258–268). Routledge. https://doi.org/10.4324/9781315742250-25

Asp, E. & Tranel, D. (2013) False tagging theory. In D. T. Stuss, & R. T. Knight (Eds.). *Principles of frontal lobe function*, 2nd ed. (pp. 383–416). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195134971.001.0001

Astuti, R. & Harris, P. (2008) Understanding mortality and the life of the ancestors in rural Madagascar. *Cognitive Science: A Multidisciplinary Journal*, *32*(4), 713–740. https://doi.org/10.1080/03640210802066907

Atran, S. (2002) *In gods we trust*. Oxford University Press.

Bail, C. A., Argyle, L. P., Brown, T. W. et al. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017) Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721.

Banas, J. A. & Rains, S. A. (2010) A meta-analysis of research on inoculation theory. *Communication Monographs*, *77*(3), 281–311.

Baron-Cohen, S. (1997) *Mindblindness: an essay on autism and theory of mind*. MIT Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition*, *21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Beecher, Henry K. (1955) The powerful placebo. *Journal of the American Medical Association* *159*(7), 1602–1606. https://doi.org/10.1001/jama.1955.02960340022006

Bendaña, J. & Mandelbaum E. (2021) The Fragmentation of Belief. In Dirk Kindermann, Cristina Borgoni, & Andrea Onofri (Eds.). *The fragmentation of mind*. Oxford University Press.

Bermúdez, J. L. (2008) *Thinking without words*. Oxford University Press.

Bingel, U., Wanigasekera, V., Wiech, K. et al. (2011) The effect of treatment expectation on drug efficacy: imaging the analgesic benefit of the opioid remifentanil. *Science translational medicine*, *3*(70), 7–14. https://doi.org/10.1126/scitranslmed.3001244

Block, N. (1980) Troubles with functionalism. *Readings in Philosophy of Psychology*, *1*, 268–305.

Bouton, M. E. (2004) Context and behavioral processes in extinction. *Learning & Memory*, *11*(5), 485–494. https://doi.org/10.1101/lm.78804

Bouton, M. E, & Bolles, R.C. (1979) Role of conditioned contextual stimuli in reinstatement of extinguished fear. *Journal of Experimental Psychology: Animal Behavior Processes*, *5*(4), 368–378. https://doi.org/10.1037//0097-7403.5.4.368

Bouton, M. E. & King, D. A. (1983) Contextual control of the extinction of conditioned fear: tests for the associative value of the context. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*(3), 248–265.

Bouton, M. E. & Ricker, S. T. (1994) Renewal of extinguished responding in a second context. *Animal Learning & Behavior*, *22*, 317–324.

Bouton, M. E., Westbrook, R. F., Corcoran, K. A., & Maren, S. (2006) Contextual and temporal modulation of extinction: behavioral and biological mechanisms. *Biological Psychiatry*, *60*(4), 352–360.

Brock, T. C. & Balloun, J. L. (1967) Behavioral receptivity to dissonant information. *Journal of personality and social psychology*, *6*(4), 413–428.

Brooks, D. C. & Bouton, M. E. (1993) A retrieval cue for extinction attenuates spontaneous recovery. *Journal of Experimental Psychology: Animal Behavior Processes*, *19*(1), 77–89.

Brownstein, M. (2015) Implicit Bias. In E. Zalta (Ed.). *The Stanford encyclopedia of philosophy*, Spring 2015.

(2018) *The implicit mind: cognitive architecture, the self, and ethics*. Oxford University Press. https://doi.org/10.1093/oso/9780190633721.001.0001

Buttelmann, D., Buttelmann, F., Carpenter, M., Call, J., & Tomasello M. (2017) Great apes distinguish true from false beliefs in an interactive helping task. *PLoS ONE*, *12*(4), 1–13, e0173793. https://doi.org/10.1371/journal.pone.0173793

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009) Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*. *112*(2), 337–342 https://doi.org/10.1016/j.cognition.2009.05.006.

Brashier, N. M. & Marsh, E. J. (2020) Judging truth. *Annual Review of Psychology*, *71*(1), 499–515. https://doi.org/10.1146/annurev-psych-010419-050807

Cheke, L. G. & Clayton, N. S. (2010) Mental time travel in animals. *WIREs Cognitive Science*, 1(6), 915–930. https://doi.org/10.1002/wcs.59

Cheyette, S. & Piantadosi, S. (2017) Knowledge transfer in a probabilistic Language Of Thought. In CogSci.

Cialdini, R. B., Trost, M. R., & Newsom, J. T. (1995) Preference for consistency: the development of a valid measure and the discovery of surprising behavioral implications. *Journal of Personality and Social Psychology*, *69*(2), 318–328. https://doi.org/10.1037/0022-3514.69.2.318

Clatterbuck, H. (2016) Darwin, Hume, Morgan, and the verae causae of psychology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *60*, 1–14. https://doi.org/10.1016/j.shpsc.2016.09.002

Clements, W. A. & Perner, J. (1994) Implicit understanding of belief. *Cognitive Development*, *9*(4), 377–395. http://dx.doi.org/10.1016/0885-2014(94)90012-4

Colloca L. & Benedetti F. (2007) Nocebo hyperalgesia: how anxiety is turned into pain. *Current Opinions in Anaesthesiology*, *20*(5),435–439. https://doi.org/10.1097/ACO.0b013e3282b972fb

Compton, J. (2020) Prophylactic versus therapeutic inoculation treatments for resistance to influence. *Communication Theory*, *30*(3), 330–343.

Cone, J., Flaharty, K., & Ferguson, M. J. (2021) The long-term effects of new evidence on implicit impressions of other people. *Psychological Science*, *32*(2), 173–188. doi:10.1177/0956797620963559

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: how, when, and why implicit evaluations can be rapidly revised. In J. M. Olson, (Ed.). *Advances in experimental social psychology: vol. 56* (pp. 131–199). Academic Press.

Davidson, D. (1982) Rational animals. *Dialectica*, *36*(4), 317–328. https://doi.org/10.1111/j.1746-8361.1982.tb01546.x

De Houwer, J., Van Dessel, P., & Moran, T. (2020) Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. *Advances in Experimental Social Psychology*, 61, 127–183.

Dennett, D. C. (1987) *The intentional stance*. MIT Press.

(1991) Real patterns. *The Journal of Philosophy*, *88*(1), 27–51. https://doi.org/10.2307/2027085

DuBrow, S., Rouhani, N., Niv, Y., & Norman, K. A. (2017) Does mental context drift or shift?. *Current Opinion in Behavioral Sciences*, *17*, 141–146.

Ecker, U. K., Lewandowsky, S., & Apai, J. (2011). Terrorists brought down the plane! – no, actually it was a technical fault: Processing corrections of emotive information. *Quarterly Journal of Experimental Psychology*, *64*(2), 283–310.

Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010) Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*(8), 1087–1100.

Egan, A. (2008) Seeing and believing: perception, belief formation and the divided mind. *Philosophical Studies*, *140*(1), 47–63. https://doi.org/10.1007/s11098–008-9225-1

Elga, A. & Rayo, A. (2021) Fragmentation and Information Access. In Dirk Kindermann, Cristina Borgoni, & Andrea Onofri (Eds.). *The fragmentation of mind*. Oxford University Press.

Fazio, L., Rand, D., & Pennycook, G. (2019) Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin & Review*, *26*, 1705–1710. https://doi.org/10.3758/s13423–019-01651-4

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015) Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*(5), 993–1002. http://dx.doi.org/10.1037/xge0000098.supp

Festinger, L., Riecken, H. W., & Schachter, S. (1956). *When prophecy fails*. University of Minnesota Press. http://dx.doi.org/10.1037/10030-000

Festinger, L., Riecken, H., & Schachter, S. (2017) *When prophecy fails: a social and psychological study of a modern group that predicted the destruction of the world*. Lulu Press, Inc.

Fodor, J. A. (1978) Propositional attitudes. *The Monist*, *61*(4), 501–523. https://doi.org/10.5840/monist197861444

(1987) *Psychosemantics: the problem of meaning in the philosophy of mind*. MIT Press.

Gendler, T. (2008) Alief and belief. *The Journal of Philosophy*, *105*(10), 634–663. https://doi.org/10.5840/jphil20081051025

Gilbert, D. T. (1991) How mental systems believe. *American Psychologist*, *46*(2), 107–119. https://doi.org/10.1037/0003-066X.46.2.107

Gilbert, D. T., Krull, D., & Malone, M. (1990). Unbelieving the unbelievable: some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*(4), 601–613. https://doi.org/10.1037/0022-3514.59.4.601

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. P. (1998) Immune neglect: a source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, *75*(3), 617–638. https://doi.org/10.1037/0022-3514.75.3.617

Greenwald, A. & Nosek, B. (2009) Attitudinal dissociation: what does it mean? In *Attitudes: insights from the new implicit measures*, R. Petty, R. Fazio, & P. Brinol (Eds.). (pp. 65–82). Psychology Press.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2014) Concepts in a probabilistic language of thought. In *The conceptual mind: new directions in the study of concepts*. MIT Press.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110–199.

Gray, H. M., Gray, K., & Wegner, D. M. (2007) Dimensions of mind perception. *Science*, *315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Gu, X., Lohrenz, T., Salas, R. et al. (2015) Belief about nicotine selectively modulates value and reward prediction error signals in smokers. *Proceedings* *of the National Academy of Sciences*, *112*(8): 2539–2544. https://doi.org/10.1073/pnas.1416639112

Gundersen, H., Specht, K., Grüner, R., Ersland, L., & Hugdahl, K. (2008) Separating the effects of alcohol and expectancy on brain activation: an fMRI working memory study. *Neuroimage*, *42*(4), 1587–1596. https://doi.org/10.1016/j.neuroimage.2008.05.037

Hampton, J. A. (1995) Testing the prototype theory of concepts. *Journal of Memory and Language*, *34*(5), 686–708. https://doi.org/10.1006/jmla.1995.1031

Hasher, L., Goldstein, D., & Toppino, T. (1977) Frequency and the conference of referential validity. *Journal of Verbal Learning & Verbal Behaviour*, *16*(1), 107–112. https://doi.org/10.1016/S0022–5371(77)80012-1.

Hawthorne, J., Rothschild, D., & Spectre, L. (2016) Belief is weak. *Philosophical Studies*, *173*(5), 1393–1404. https://doi.org/10.1007/s11098–015-0553-7

van Holst, R. J., Clark, L., Veltman, D. J., van den Brink, W., & Goudriaan, A. E. (2014) Enhanced striatal responses during expectancy coding in alcohol dependence. *Drug and Alcohol Dependence*, *142*, 204–208. https://doi.org/10.1016/j.drugalcdep.2014.06.019

Howard, S., Avarguès-Weber, A., Garcia, J., Greentree A.D., & Dyer. A. D. (2019) Numerical ordering of zero in honeybees. *Science*, *360*(6393), 1124–1126. https://doi.org/10.1126/science.aar4975

Hughes, S., Ye, Y., Van Dessel, P., & De Houwer, J. (2019) When people co-occur with good or bad events: graded effects of relational qualifiers on evaluative conditioning. *Personality and Social Psychology Bulletin*, *45*(2), 196–208. https://doi.org/10.1177/0146167218781340

Icard, T. (2016) Subjective probability as sampling propensity. *Review of Philosophy and Psychology*, *7*(4), 863–903.

Jacobs, K. W. & Nordan, F. M. (1979) Classification of placebo drugs: effect of color. *Perceptual and motor skills*, *49*(2), 367–372. https://doi.org/10.2466/pms.1979.49.2.367

Jensen, G., Alkan, Y., Ferrera, V. P., & Terrace, H. S. (2018) Reward associations do not explain transitive inference performance in monkeys. *Science Advances*, *5*(7), Article aaw2089. https://doi.org:10.7287/peerj.preprints.26889v1

Katz, Y., Goodman, N.D., Kersting, K., Kemp, C., & Tenenbaum, J.B. (2008). Modeling semantic cognition as logical dimensionality reduction. *Proceedings of the Cognitive Science Society, 30*.

Kemp, C., Goodman, N., & Tenenbaum, J. (2008). Learning and using relational theories. *Advances in Neural Information Processing Systems*, *20*, 753–760.

Knowles, E. S. & Condon, C. A. (1999) Why people say "yes": a dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, *77*(2), 379–386. https://doi.org/10.1037/0022-3514.77.2.379

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016) Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114. https://doi.org/10.1126/science.aaf8110

Kufahl P., Li, Z., Risinger, R. et al. (2008) Expectation modulates human brain responses to acute cocaine: a functional magnetic resonance imaging study. *Biological Psychiatry*, *63*(2), 222–230. https://doi.org/10.1016/j.biopsych.2007.03.021

Kumkale, G. T. & Albarracín, D. (2004) The sleeper effect in persuasion: a meta-analytic review. *Psychological Bulletin*, 130(1), 143–172. https://doi.org/10.1037/0033-2909.130.1.143

Kurdi, B. & Banaji, M. R. (2017) Repeated evaluative pairings and evaluative statements: how effectively do they shift implicit attitudes? *Journal of Experimental Psychology: General*, *146*(2), 194–213. http://dx.doi.org/10.1037/xge0000239

(2019) Attitude change via repeated evaluative pairings versus evaluative statements: shared and unique features. *Journal of Personality and Social Psychology*, *116*(5), 681–703. http://dx.doi.org/10.1037/pspa0000151

Kurdi, B. & Dunham, Y. (2020) Propositional accounts of implicit evaluation: taking stock and looking ahead. *Social Cognition*, *38*(Supplement), s42–s67. http://doi.org/10.1521/soco.2020.38.supp.s42

(ms) Sensitivity of implicit evaluations to accurate and erroneous propositional inferences.

Lai, C. K., Skinner, A. L., Cooley, E. et al. (2016) Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016. https://psycnet.apa.org/doi/10.1037/xge0000179

Lacassagne, D., Béna, J., & Corneille, O. (2022) Is Earth a perfect square? Repetition increases the perceived truth of highly implausible statements. *Cognition*, 223, 105052. https://doi.org/10.1016/j.cognition.2022.105052

Lange, M. (2002) Who's Afraid of Ceteris-Paribus Laws? Or: How I Learned to Stop Worrying and Love Them. *Erkenntnis* 57, 407–423. https://doi.org/10.1023/A:1021546731582

Lewis, D. (1980) Mad pain and Martian pain. In Ned Block (Ed.). *Readings in the philosophy of psychology* (pp. 216 S.222). Harvard University Press.

Liang, P., Jordan, M. I., & Klein, D. (2010) Learning programs: A hierarchical Bayesian approach. In Proceedings of the 27th International Conference on Machine Learning (ICML-10) (pp. 639–646).

Lurz, R. W. (2015) *Mindreading animals: the debate over what animals know about other minds*. MIT Press. https://doi.org/10.7551/mitpress/9780262016056.001.0001

Mandelbaum, E. (2013) Against alief. *Philosophical Studies*, *165*(1), 197–211. https://doi.org/10.1007/s11098–012-9930-7

(2014) Thinking is believing. *Inquiry*, *57*(1), 55–96. https://doi.org/10.1080/0020174X.2014.858417

(2015a) Associationist theories of thought. In Edward N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition. https://plato.stanford.edu/archives/sum2017/entries/associationist-thought/

(2015b) The automatic and the ballistic: modularity beyond perceptual processes. *Philosophical Psychology*, *28*(8), 1147–1156. https://www.doi.org/10.1080/09515089.2014.950217

(2016) Attitude, inference, association: on the propositional structure of implicit bias. *Noûs*, *50*(3), 629–658. https://doi.org/10.1111/nous.12089

(2018) Seeing and conceptualizing: modularity and the shallow contents of perception. *Philosophy and Phenomenological Research*, *97*(2), 267–283. https://doi.org/10.1111/phpr.12368

(2019) Troubles with Bayesianism: an introduction to the psychological immune system. *Mind & Language*, *34*(2), 141–157. https://doi.org/10.1111/mila.12205

(Forthcoming) *A psychofunctional theory of belief*. Oxford University Press.

Mandelbaum, E. & Quilty-Dunn, J. (2015). Believing without reason or: why liberals shouldn't watch Fox News. *The Harvard Review of Philosophy*, *22*, 42–52. https://doi.org/10.5840/harvardreview2015226

Mann, T. C. & Ferguson, M. J. (2015) Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology*, *108*(6), 823–849. https://doi.org/10.1037/pspa0000021

(2017). Reversing implicit first impressions through reinterpretation after a two-day delay. *Journal of Experimental Social Psychology*, *68*, 122–127. https://doi.org/10.1016/j.jesp.2016.06.004

Mann T. C., Kurdi B., & Banaji M. R. (2019) How effectively can implicit evaluations be updated? Using evaluative statements after aversive repeated evaluative pairings. *Journal of Experimental Psychology: General*. Advance Online Publication. https://doi.org/10.1037/xge0000701.

McHoskey, J. W. (1995) Case closed? On the John F. Kennedy assassination: biased assimilation of evidence and attitude polarization. *Basic and Applied Social Psychology*, *17*(3), 395–409. https://doi.org/10.1207/s15324834basp1703_7

McGuire W. J. (1964). Inducing resistance to persuasion: some contemporary approaches. In L. Berkowitz (Ed.). *Advances in Experimental Social Psychology*, *Vol. 1* (pp. 191–229). Academic Press.

Mercier, H. (2020) *Not born yesterday: the science of who we trust and what we believe*. Princeton University Press.

Mollica, F. & Piantadosi, S. (2015). Towards semantically rich and recursive word learning models. *Proceedings of the Cognitive Science Conference*, *37*, 1607–1612.

Nichols, S. (2021) *Rational rules: towards a theory of moral learning*. Oxford University Press.

Nielsen, M. & Stewart, R. T. (2020) Persistent disagreement and polarization in a Bayesian setting. *The British Journal for the Philosophy of Science*, *72* (1), 51–78.

Núñez, R., Allen, M., Gao, R., Rigoli, C. M., Relaford-Doyle, J., & Semenuks, A. (2019) What happened to cognitive science?. *Nature Human Behaviour*, *3* (8), 782–791. https://doi.org/10.1038/s41562–019-0626-2

Onishi, K. H. & Baillargeon, R. (2005) Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621

van Orman Quine, W. & Ullian, J. (1978) *The web of belief*. Random House.

Ossipov, M. H., Dussor G. O., & Porreca, F. (2010) Central modulation of pain. *Journal of Clinical Investigation*, *120*(11), 3779–3787. https://doi.org/10.1172/JCI43766

Pennycook, G., Cannon, Tyrone D. & Rand, David, G. (2018) Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://www.doi.org/10.1037/xge0000465

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015) On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, *10*(6), 549–563.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021) Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*, 1–6.

Pennycook, G. & Rand, D. G. (2019) Lazy, not biased: susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

(2021) The Psychology of Fake News. *Trends in Cognitive Sciences*, 25(5), 388–402.

Pepperberg, I. M., Gray, S. L., Mody, S., Cornero, F. M., & Carey, S. (2019) Logical reasoning by a grey parrot? A case study of the disjunctive syllogism. *Behaviour*, *156*(5–8), 409–445. https://doi.org/10.1163/1568539x-00003528

Petrocelli, J. V., Tormala, Z. L., & Rucker, D. D. (2007) Unpacking attitude certainty: attitude clarity and attitude correctness. *Journal of Personality and Social Psychology*, *92*(1), 30–41. https://doi.org/10.1037/0022-3514.92.1.30

Phillips, J., Buckwalter, W., Cushman, F. et al. (2020) Knowledge before belief. *Behavioral and Brain Science*, 8(44), E140. https://doi.org/10.1017/S0140525X20000618

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: a formal model of numerical concept learning. *Cognition 123*(2), 199–217.

(2016) The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, 123(4), 392–424. https://doi.org/10.1037/a0039980

Porot, N. J. (2019) Some non-human languages of thought. (Unpublished doctoral dissertation). City University of New York, Graduate Center.

Povinelli, D. J. & Vonk, J. (2003) Chimpanzee minds: suspiciously human? *Trends in Cognitive Sciences*, *7*(4), 157–160. https://doi.org/10.1016/s1364-6613(03)00053-6

Quilty-Dunn, J. & Mandelbaum, E. (2018a) Against dispositionalism: belief in cognitive science. *Philosophical Studies*, *175*(9), 2353–2372. https://doi.org/10.1007/s11098–017-0962-x

(2018b) Inferential transitions. *Australasian Journal of Philosophy*, *96*(3), 532–547. https://doi.org/10.1080/00048402.2017.1358754

Rescorla, M. (2016) Bayesian sensorimotor psychology. *Mind & Language*, *31*(1), 3–36.

Ritchie, K. (2016) Can semantics guide ontology? *Australasian Journal of Philosophy*, *94*(1), 24–41. https://doi.org/10.1080/00048402.2015.1045912

Rosati, A. G. & Santos, L. R. (2016) Spontaneous metacognition in rhesus monkeys. *Psychological Science*, *27*(9), 1181–1191. https://doi.org/10.1177/0956797616653737

Ross, L. (1977) The intuitive psychologist and his shortcomings: distortions in the attribution process. *Advances in Experimental Social Psychology*, *10*, 173–220. https://doi.org/10.1016/S0065–2601(08)60357-3

Schachter, S. & Singer, J. E. (1962) Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379–399. https://doi.org/10.1037/h0046234

Schwitzgebel, E. (2013) A dispositional approach to attitudes: thinking outside of the belief box. In N. Nottelmann (Ed.). *New essays on belief* (pp. 75–99). Palgrave Macmillan. https://doi.org/10.1057/9781137026521_5

Seifert, C. (2002) The Continued Influence of Misinformation in Memory: What Makes a Correction Effective? In B. Ross (Ed.). *Psychology of Learning and Motivation*, vol. 41 (pp. 265–292). Academic Press.

Singer, D. J., Bramson, A., Grim, P. et al. (2019) Rational social and political polarization. *Philosophical Studies*, *176*(9), 2243–2267.

Skurnik, I., Yoon, C., Park, D. C., & Schwarz, N. (2005) How warnings about false claims become recommendations. *Journal of Consumer Research*, *31*(4), 713–724.

Smith, J. D., Couchman, J. J., & Beran, M. J. (2014). The highs and lows of theoretical interpretation in animal-metacognition research. *The Cognitive Neuroscience of Metacognition*, 367(1594), 1297–1309. https://doi.org/10.1007/978-3-642-45190-4_5

Sperber, D. (1997) Intuitive and reflective beliefs. *Mind and Language*, *12*(1), 67–83.

Sperber, D., Clément, F., Heintz, C. et al. (2010) Epistemic vigilance. *Mind and Language*, *25* (4), 359–393.

Stich, S. P. (1979) Do animals have beliefs? *Australasian Journal of Philosophy*, *57* (1), 15–28. https://doi.org/10.1080/00048407912341011

Storms, M. D. & Nisbett, R. E. (1970). Insomnia and the attribution process. *Journal of Personality and Social Psychology*, *16*(2), 319–328. https://doi.org/10.1037/h0029835

Tibbetts, E., Agudelo, J., Pandit, S., & Riojas, J. (2019) Transitive inference in Polistes paper wasps. *Biology Letters*, 15(5), Article 20190015. https://doi.org/10.1098/rsbl.2019.0015

Tracey, I. (2010) Getting the pain you expect: mechanisms of placebo, nocebo and reappraisal effects in humans. *Nature Medicine*, *16*(11), 1277–1283. https://doi.org/10.1038/nm.2229

Tracy, R. Young, S. Porot, N., & Mandelbaum, E. (under review) Disfluency Attenuates the Reception of Pseudoprofound and Postmodernist Bullshit.

Trouche, E., Johansson, P., Hall, L., & Mercier, H. (2016) The selective laziness of reasoning. *Cognitive Science 40*(8), 2122–2136.

Unkelbach, C. & Rom, S. C. (2017) A referential theory of the repetition-induced truth effect. *Cognition*, *160*, 110–126. https://doi.org/10.1016/j.cognition.2016.12.016

Van Dessel, P., De Houwer, J., Gast, A., Smith, C. T., & De Schryver, M. (2016) Instructing implicit processes: when instructions to approach or avoid influence implicit but not explicit evaluation. *Journal of Experimental Social Psychology*, *63*, 1–9. https://doi.org/10.1016/j.jesp.2015.11.002

Van Dessel, P., Ye, Y., & De Houwer, J. (2019) Changing deep-rooted implicit evaluation in the blink of an eye: negative verbal information shifts automatic liking of Gandhi. *Social Psychological and Personality Science*, *10*(2), 266–273. https://doi.org/10.1177/1948550617752064

Van Leeuwen, N. (2014) Religious credence is not factual belief. *Cognition*, *133* (3), 698–715. https://doi.org/10.1016/j.cognition.2014.08.015

(2017) Two paradigms for religious representation: the physicist and the playground (a reply to Levy). *Cognition*, *164*, 206–211. https://doi.org/10.1016/j.cognition.2017.03.021

Vogel, T., Silva, R. R., Thomas, A., & Wänke, M. (2020) Truth is in the mind, but beauty is in the eye: fluency effects are moderated by a match between fluency source and judgment dimension. *Journal of Experimental Psychology: General*, *149*(8), 1587–1596. doi:10.1037/xge0000731

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal Decisions from Very Few Samples. *Cognitive Science*, *38*(4), 599–637. https://doi.org/10.1111/cogs.12101

Vul, E. & Pashler, H. (2008) Measuring the crowd within: probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647. https://doi.org/10.1111/j.1467-9280.2008.02136.x

Waber, R. L., Shiv, B., Carmon, Z., & Ariely, D. (2008) Commercial features of placebo and therapeutic. *Journal of the American Medical Association*, *299*(9), 1016–1017. https://doi.org/10.1001/jama.299.9.1016

Watson, A., Power, A., Brown, C., El-Deredy, W., & Jones, A. (2012) Placebo analgesia: cognitive influences on therapeutic outcome. *Arthritis Research & Therapy*, *14*, Article 206. https://doi.org/10.1186/ar3783.

Weisberg, J. (2020) Belief in psyontology. *Philosophers' Imprint*, *20*(11), 1–27.

Webster, D. M. & Kruglanski, A. W. (1994) Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, *67*(6), 1049–1062. https://doi.org/10.1037//0022-3514.67.6.1049

Wegner, D. M., Coulton, G. F., & Wenzlaff, R. (1985) The transparency of denial: briefing in the debriefing paradigm. *Journal of Personality and Social Psychology*, *49*(2), 338–346.

Whissel, C., Abramson, C. I., & Barber, K. I. (2013) The search for cognitive terminology: an analysis of comparative psychology journal titles. *Behavioral Sciences*, *3*(1), 133–142.

Wiklund, R. A. & Rosenbaum, S. H. (1997) Anesthesiology (part one of two). *New England Journal of Medicine*, *337*(17), 1215–1219. https://doi.org/10.1056/nejm199710233371707

Zanna, M. & Cooper, J. (1974) Dissonance and the pill: an attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology*, *29*(5), 703–709. https://doi.org/10.1037/h0036651

Zettlemoyer, L. S. & Collins, M. (2005). Learning to map sentences to logical form: Structured classification with probabilistic categorical grammars. Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence, pp. 658–666.