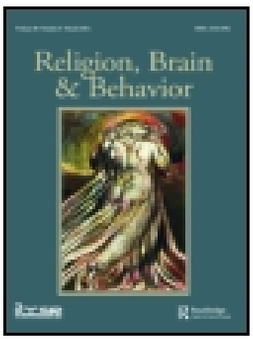

Routledge
Taylor & Francis Group

Religion, Brain & Behavior

ISSN: 2153-599X (Print) 2153-5981 (Online) Journal homepage: http://www.tandfonline.com/loi/rrbb20

Explaining agency detection within a domain-specific, culturally attuned model

Joni Y. Sasaki & Adam S. Cohen

To cite this article: Joni Y. Sasaki & Adam S. Cohen (2017): Explaining agency detection within a domain-specific, culturally attuned model, Religion, Brain & Behavior, DOI: 10.1080/2153599X.2017.1387592

To link to this article: http://dx.doi.org/10.1080/2153599X.2017.1387592

Published online: 17 Nov 2017.

RELIGION, BRAIN & BEHAVIOR, 2017
https://doi.org/10.1080/2153599X.2017.1387170

Routledge
Taylor & Francis Group

TARGET ARTICLE

Predictive coding in agency detection

Marc Andersen[a,b,c]

[a]Department of Culture and Society, Aarhus University, Aarhus, Denmark; [b]Religion, Cognition and Culture, Aarhus University, Aarhus, Denmark; [c]Interacting Minds Centre, Aarhus University, Aarhus, Denmark

ABSTRACT

Agency detection is a central concept in the cognitive science of religion (CSR). Experimental studies, however, have so far failed to lend support to some of the most common predictions that follow from current theories on agency detection. In this article, I argue that predictive coding, a highly promising new framework for understanding perception and action, may solve pending theoretical inconsistencies in agency detection research, account for the puzzling experimental findings mentioned above, and provide hypotheses for future experimental testing. Predictive coding explains how the brain, unbeknownst to consciousness, engages in sophisticated Bayesian statistics in an effort to constantly predict the hidden causes of sensory input. My fundamental argument is that most false positives in agency detection can be seen as the result of top-down interference in a Bayesian system generating high prior probabilities in the face of unreliable stimuli, and that such a system can better account for the experimental evidence than previous accounts of a dedicated agency detection system. Finally, I argue that adopting predictive coding as a theoretical framework has radical implications for the effects of culture on the detection of supernatural agency and a range of other religious and spiritual perceptual phenomena.

ARTICLE HISTORY
Received 26 April 2016
Accepted 24 January 2017

KEYWORDS
Agency detection; HADD; predictive coding; perception; supernatural agents; religion; epidemiology; cognitive science of religion

1. Introduction

Although the field was anticipated by Dan Sperber as early as 1975, it is generally argued that the cognitive science of religion (CSR) proper was born in 1980 when Stewart Guthrie proposed the first cognitive theory of religion. In what is now considered a classic article, Guthrie proposed that religious belief can be explained as the by-product of an evolved perceptual bias in human cognition whereby ambiguous non-agentive phenomena are frequently misperceived as human-like agents (Guthrie, 1980). During the second of the two decades subsequent to the publication of Guthrie's paper, CSR established itself as an independent subfield with a primary focus on three issues, all of which were heavily informed by evolutionary psychology: (1) religious rituals (Lawson & McCauley, 1990; Whitehouse, 1992); (2) cognitive constraints in the representation of supernatural agents (Barrett, 1999; Barrett & Keil, 1996; Boyer, 1996); and (3) the acquisition of supernatural concepts (Barrett, 2000; Boyer, 1994; Guthrie, 1993; Guthrie, 1980). Today, CSR encompasses a vast range of topics (for an overview, see Barrett, 2011), but agency detection remains a central concept in the field.

The fundamental argument of all literature on agency detection is that humans have evolved a perceptual apparatus that is hard-wired to be overly sensitive to the detection of agents.[1] This means that our perceptual system will have a tendency to produce false positives, such that humans

CONTACT Marc Andersen mana@cas.au.dk

will detect agents in the environment when none are actually present (Atran, 2002; Barrett, 2000; Guthrie, 1993). Over the 35 years since its inception, research on agency detection has not only pointed to the theoretical importance of oversensitive agency detection as a factor in the generation and maintenance of religious belief, but has also given rise to numerous experimental studies on the subject. These studies, however, have thus far failed to lend support to some of the most frequent predictions that follow from the theories. This may be due to the fact that current theories on agency detection in CSR are the legacy of a certain type of evolutionary psychology in which the human mind was thought to consist of a collection of computationally distinct and domain-specific "modules."

Today, exciting new developments in cognitive science are encouraging a stronger focus on processual and domain-general models of perception and cognition. One such model currently taking firm hold in neuroscience is predictive coding. Predictive coding states that the brain is an expectation-reliant prediction machine that constantly anticipates the incoming sensory information it is about to encounter. Failed predictions generate "prediction errors" that are used to generate better predictions or facilitate the slower process of learning (Clark, 2016). Predictive coding not only has unifying power and great potential in terms of explanatory scope; it is also backed by powerful theoretical arguments and mounting empirical evidence (Bubic, Von Cramon, & Schubotz, 2010; Clark, 2013, 2016; Den Ouden, Kok, & De Lange, 2012; Friston & Stephan, 2007; Hohwy, 2013, 2014; Huang & Rao, 2011). Predictive coding is also starting to be integrated into CSR (Andersen, Schjoedt, Nielbo, & Sørensen, 2014; Hermans, 2015; Nielbo & Sørensen, 2013; Schjoedt et al., 2013a, 2013b; Taves & Asprem, 2017), but has not yet been applied to research on agency detection. In this article, I argue that predictive coding may solve pending theoretical inconsistencies in agency detection research, account for puzzling experimental findings, and provide hypotheses for future experimental testing. I further argue that adopting predictive coding as a theoretical framework radically changes our understanding of agency detection and a range of other religious and spiritual perceptual phenomena by providing greater room for the incorporation of cultural variation in the tendency to interpret ambiguous perceptual data as (supernatural) agents.

2. Agency detection

Most people have experienced spotting someone or something animate when nobody was actually there. Be it a rattling in the bushes taken for an animal or a bump in the night taken for a burglar, such mistakes in agency detection have long been a subject of interest for CSR researchers. The idea that supernatural agent beliefs are somehow caused or affected by mistakes in the human agency detection system was initially proposed by Stewart Guthrie (1980). In his writings, Guthrie argues that beliefs in supernatural agents arise from perceptual (and, latterly, conceptual) errors that humans are hard-wired to frequently commit (Guthrie, 1980; Guthrie, 1993, 1996, 2002, forthcoming). These mistakes or false positives, Guthrie claims, will inevitably arise because humans generally have a low threshold for judging when we have detected an agent. This low threshold for agency detection Guthrie interprets as a result of our evolutionary past: "This strategy has evolved, based on a good principle: Better safe than sorry" (Guthrie, 2013, p. 261). Guthrie's account has inspired a wide range of research and spurred two similar theoretical accounts of the relationship between the inner workings of the human agency detection system and religious belief, namely those put forth by psychologist Justin Barrett and anthropologist Scott Atran. These authors generally agree with Guthrie's claim that false positives produced by the human agency detection system are important with respect to belief in supernatural agents (Atran, 2002; Atran & Norenzayan, 2004; Barrett, 2000, 2004; Barrett & Lanman, 2008). Their accounts differ, however, in respect to (1) *when* the human agency detection system is particularly prone to produce false positives, (2) *what* effect it is thought to have on religious belief, and (3) *what* the specified cognitive model looks like.

(1) Evidently, humans and the vast majority of other mammals excel at spotting agents in their nearby environment. It takes but a fraction of a second to process whether agents are nearby, who

those same agents might be, and how many of them there are. The primary question of interest, however, is under what conditions does the human agency detection system activate inappropriately? Collectively, Guthrie, Barrett, and Atran refer to a range of experimental studies demonstrating that children and adults watching moving dots or geometrical forms on a screen routinely interpret these as interacting agents with particular goals and motivations (Bloom & Veres, 1999; Csibra, Gergely, Bíró, Koós, & Brockbank, 1999; Heider & Simmel, 1944; Premack & Premack, 1995; Rochat, Morgan, & Carpenter, 1997). These experiments suggest that, from infancy, we are very sensitive to movement that looks as if it is goal-directed or self-propelled, and they are typically taken as evidence demonstrating a tendency to swiftly attribute agency to even the simplest non-living stimuli, even under conditions where we are reflectively aware that this is not really the case. Adding to these findings, Atran argues that humans are more prone to over-detect agents under uncertain and dangerous circumstances and in situations of opportunity:

> A cognitive schema for recognizing and interpreting animate agents may be a crucial part of our evolutionary heritage, which primes us to anticipate intention in the unseen causes of uncertain situations that carry the risk of danger or the promise of opportunity, such as predators, protectors, and prey. (Atran, 2002, p. 61)

Guthrie, on the other hand, argues that all phenomena that humans encounter during their lives are ambiguous in essence and in need of interpretation in order to be understood (Guthrie, 2013; Guthrie, 1980). Humans, he claims, have evolved a perceptual default to see the world "neither as what we 'want' to see nor as what is most likely, but as what matters most" (Guthrie, 1996, p. 418). Our frequent experience of false positives results from the sheer importance to us of interactions with our fellow human beings and other intentional agents: "Perception is betting. … In this betting, we choose the most significant possibilities we know: the spider, not the thread; the burglar, not the wind" (Guthrie, 2013, p. 261). Thus, Guthrie points to a more general perceptual bias of applying human-like models to phenomena in the world, which is activated relatively equally across most times and contexts (Guthrie, 1993).[2] Finally, Barrett, who dubbed this supposed cognitive module the "Hypersensitive Agency Detection Device" (HADD) (Barrett, 2000), argues that it may also be affected by personal context, histories, and dispositions. In line with Atran, Barrett asserts that the activity of the HADD may increase during situations of urgency that "suggest survival to our prehistorically created minds" (Barrett, 2004, p. 39), meaning situations where humans are either on the lookout for prey or under the impression that they might themselves be the object of predation (Barrett, 2004). All three agree, however, that the human agency detection system is particularly prone to produce false positives in the face of ambiguous stimuli that can be interpreted in a multitude of ways (Atran, 2002; Barrett, 2004; Guthrie, 1993).

(2) The accounts set forth by Guthrie, Atran, and Barrett differ in terms of what effects the human agency detection system is thought to have on religious belief. Atran argues that the human cognitive proclivity to detect agents under conditions of uncertainty facilitates the emergence of malevolent deities in all cultures, just as the human proclivity to attach to caregivers facilitates the emergence of benevolent deities (Atran, 2002; Atran & Norenzayan, 2004). Guthrie, on the other hand, argues more boldly that religion as such can be explained as a by-product of the perceptual biases produced by the human agency detection system (Guthrie, 1993, 2002; Guthrie, 1980). Some scholars have subsequently raised doubts about the supposed centrality of the human agency detection system in generating beliefs in supernatural agents (Atran, 2002; Boyer, 2001; Gervais & Henrich, 2010; Gervais, Willard, Norenzayan, & Henrich, 2011), arguing that the false positives produced by the system are often quickly ignored, easily overridden, and seldom transform into stable beliefs about supernatural beings. In response, Barrett and Lanman have argued that, while false positives from the agency detection system may not be directly responsible for belief in supernatural agents, they may well serve to encourage, strengthen, and spread such supernatural agent beliefs as may already be in place within a given cultural context (Barrett, 2007; Barrett & Lanman, 2008).

(3) Guthrie, Barrett, and Atran all agree that the ultimate cause of the human agency detection system's tendency to produce false positives is that evolution shaped the human cognitive system

to react in accordance with the mantra "better safe than sorry" (Atran, 2002; Barrett, 2004; Guthrie, 1993). As Guthrie puts it: "Walking in the woods, it's better to mistake a stick for a snake, or a boulder for a bear, than the reverse. If we're right, we gain much, and if wrong, we lose little" (Guthrie, 2013, p. 261). The accounts differ, however, in terms of the proximate model of cognition posited. In this regard, Guthrie sets forth the most elaborate and detailed model of agency detection. He argues that humans understand and interpret the world through the application of cognitive schemata/models, an argument that is fairly uncontroversial. The salient point in his account, however, is that humans will always try to fit an agent model to a phenomenon before other models are tried, because agents are what ultimately matter most to us (Guthrie, 1993; Guthrie, 1980). This, he claims, is the primary cause of the human tendency to attribute agency to non-living things; because of the vast host of events that could in principle have been caused by humans, a human model will often fit non-human phenomena. The cognitive models proposed by Barrett and Atran, however, are both either vague or underspecified in terms of proximate mechanisms. Atran argues that the human agency detection system is the product of a cognitive module evolved to handle the *proper domain* of information signifying the presence of agents. Its *actual domain*, however, by which he refers to *all* information that satisfies the input conditions of the module, extends further, which is why humans tend to produce false positives in agency detection (Atran, 2002). Although Atran clarifies *what* the module reacts to (e.g., self-propelled movement), he does not specify *how* the cognitive module achieves this. This makes Atran's proximate model something of a "black-box" explanation, where the actual inner workings are left unspecified. The same is true for the proximate model set forth by Barrett. While Barrett provides a detailed account of what stimuli the HADD will react to and under what circumstances it is particularly prone to error, the proximate cognitive model of *how* the HADD does this is left unspecified.

The lack of explicit proximate models in agency detection research is a non-trivial matter. Without a clear and explicit cognitive model of the human agency detection system, it is difficult to make predictions about when in particular the system should be hyperactive and prone to produce false positives. How exactly, for example, has the HADD on the one hand evolved to be so hyperactive that it gives rise to the detection of non-existent agents in the surrounding environment while on the other remaining well calibrated enough to facilitate precise group coordination activities such as large game hunting (Green, 2015)? Though selection pressures may sacrifice absolute perceptual accuracy to increased utility in pursuit of genetic fitness, there must be a point where a cognitive system *overly* prone to making false inferences should not be favored by natural selection (Boyer, 2001). A more robustly elaborated proximate model is therefore necessary in order to specify precise factors that moderate the misperception of agency. Similarly, is it the case that humans have a general perceptual bias of applying human-like models to phenomena in the world that is activated equally across most times and contexts (Guthrie, 1993), or are there strong cultural, dispositional, and situational determinants of the hyperactivity of the HADD (Atran, 2002; Barrett, 2004)? To solve theoretical inconsistencies such as these, and formulate clear hypotheses for experimental testing, the field would greatly benefit from the application of clear proximate models from cognitive neuroscience.

2.1. Experimental evidence

Despite this lack of clearly formulated proximate models, a number of experimental studies have tried to shed light on the factors underlying the hypothesized hyperactivity of the HADD and have, taken as a whole, found mixed results. In one study, it was found that paranormal and religious believers experience more "false alarms" compared to skeptics when asked to detect human-like faces in artifacts or scenery (Riekki, Lindeman, Aleneff, Halme, & Nuortimo, 2013). Contrary to this finding, another study found that paranormal believers were more accurate than skeptics in detecting faces in a face-house categorization task (van Elk, 2015). A separate study found that paranormal believers were more likely than skeptics to detect agency in a biological motion perception task

under low and intermediate levels of noise, but this effect disappeared at high noise levels (van Elk, 2013). Finally, van Elk and colleagues conducted a series of five priming experiments investigating the effect of semantic supernatural agent primes on agency detection. Using biological motion tasks and face-house categorization tasks under different noise levels, they found that human primes facilitated agency detection but supernatural agent primes did not. In this case, the researchers found an effect of religiosity in agency detection in three experiments but no such effect in two others. Interestingly, however, no general response bias towards agent detection as a function of heightened noise levels was found in this series of experiments (van Elk, Rutjens, van der Pligt, & van Harrevel, 2016).

Taken as a whole, these experimental findings can hardly be said to support current theoretical models of the HADD. If humans process ambiguous information according to the mantra "better safe than sorry," we would expect a response bias towards agent categorization as a function of noise levels in classical paradigms such as the face-house categorization task and the biological motion task, but no such tendency has been found (van Elk, 2015; van Elk et al., 2016). These are surprising findings indeed when placed alongside the standard claim that the HADD becomes hyperactive when encountering ambiguous stimuli (Atran, 2002; Barrett, 2004; Guthrie, 1993). Similarly, Guthrie's importance-driven model of agency detection would predict a general response bias in these paradigms towards categorizing stimuli as agents, but no such bias is present in the empirical findings (van Elk, 2015; van Elk et al., 2016). To the author's knowledge, the only empirical work that somewhat supports current theoretical models of the HADD is correlational evidence from a survey study which found that experiences of nonreligious supernatural agency were more likely to have occurred in threatening and ambiguous environments than spiritual, religious, supernatural, or paranormal experiences (Barnes & Gibson, 2013). Finally, semantic supernatural agent primes have so far been shown not to facilitate agency detection, while experimental findings regarding the effects of cultural and situational determinants on agency detection remain mixed.

A related field of empirical research that may have some bearing on the question of agency detection is that concerned with the cognitive and neural components involved in the feeling of *sensed presence*. Sensed or felt presence refers to the sensation that someone or something animate is present in the nearby environment of an individual. Sensed presence can manifest itself in a variety of ways, and is often experienced despite the absence of any identifiable stimuli corresponding to the experience (Solomonova, Frantova, & Nielsen, 2011). Sensed presence has been documented extensively in research on sleep paralysis and has been linked to the "old hag phenomenon" whereby people suffer the experience of waking up at night, paralyzed, and gripped by the distressing feeling that "someone is there" (Hufford, 1982; Solomonova et al., 2011; Taves, 2009).

Within this field, cognitive neuroscientist Michael Persinger has claimed that the application of weak complex magnetic fields to the temporal lobes can elicit the sensed presence of a sentient being. This sensation of presence, he claims, is the prototype of a vast swathe of mystical, religious, and paranormal experiences, including spirit visitations, alien abductions, and the experience of gods (Persinger, 1983, 2001, 2002; Persinger, Saroka, Koren, & St-Pierre, 2010). The proximate model proposed by Persinger specifies that the brain contains a sense of self in each of its hemispheres. Under normal circumstances the left hemisphere's sense of self is dominant, and supplies individuals with their everyday sense of self. However, by stimulating the right hemisphere's temporal lobe, Persinger claims, the right hemisphere's sense of self comes to intrude on that of the left, thereby creating the unique experience of someone or something being present (Persinger, 2001; Persinger et al., 1994, 2010). In his earlier writings Persinger argues that all cultures contain techniques that facilitate mystical and religious experiences through temporal lobe activation of this sort (Persinger, 1983), and he extends this assertion in later writings to the claim that all modern societies are filled with electromagnetic devices that can influence the activity of the temporal lobe via magnetic stimulation (Persinger et al., 2010). For instance, Persinger makes reference to a case report of a woman reporting visions of an apparition at night, but is quick to localize the culprit as the woman's alarm clock standing next to her bedside (Persinger, 2001). Of course, studies like these are correlational, and cannot be used to infer causality. For the purpose of experimentally inducing the experience of a

sensed presence and testing his causal hypothesis, Persinger uses an electromagnetic device, the so-called God helmet, which stimulates the temporal lobes of participants with weak complex magnetic fields while participants are placed in sensory deprivation (Hill & Persinger, 2003). Persinger claims that this setup enables him to induce an experience of sensed presence in 80% of the general population (Persinger, in Granqvist et al., 2005). Persinger's God helmet studies, however, are controversial for a number of methodological and theoretical reasons (Granqvist et al., 2005; Schjoedt, 2009). First, the theoretical framework on sense of self that Persinger employs is not supported in conventional theories of cognitive neuroscience (e.g., Hohwy, 2007). As other researchers have argued, experimental neuroscience on religious experience should rely on conventional theories of brain function and cognitive processing rather than developing new controversial models of cognition (Bulbulia & Schjoedt, 2013; Schjoedt, 2009). Second, while conventional transcranial magnetic stimulation (TMS) is a well-established technique for eliciting brain activity, the magnetic fields emitted from Persinger's God helmet are approximately one million times weaker (Persinger, 2002), approximating the magnetic field emitted by a standard hairdryer. From a theoretical perspective, it would seem unlikely that such weak magnetic fields could produce currents strong enough to depolarize neurons (Ruohonen, 1998), although Persinger claims that the specific signature of the waveform is essential for attaining the effect (Persinger, in Granqvist et al., 2005). Third, in an effort to replicate Persinger's results, Pehr Granqvist and colleagues conducted a randomized double-blinded study to test the effects of the God helmet's magnetic stimulus. After borrowing the equipment from Persinger's own lab, the Swedish research team employed two conditions, one in which the helmet was switched on and one in which it was switched off, and found no effect of the magnetic stimulus (Granqvist et al., 2005). Some participants reported unusual experiences, but these were evenly distributed between the target and control groups. Interestingly, reported experiences were predicted by participants' suggestibility scores, and the interpretive frameworks applied by individuals to the experiences in question were predicted by participants' reported levels of religiosity (Granqvist et al., 2005; Granqvist & Larsson, 2006). These findings indicate that it is not the magnetic fields of the God helmet itself, but the *idea* of the God helmet and the context in which it is administered that drive these spectacular effects (Granqvist et al., 2005).[3] Thus, a confluence of individual expectation, contextual factors, individual levels of suggestibility, and sensory deprivation seems able to facilitate experiences of the presence of an agent, despite such an agent's actual absence.

2.2. Summary

Research on agency detection is permeated by the idea that humans have a hypersensitive agency detection device with a low threshold for judging when it has detected an agent. This threshold is thought to be a direct result of our evolutionary past, with the principle "better safe than sorry" postulated to have been the guiding mantra for the development of the system. From this, it has generally been claimed that the system is particularly prone to error upon encountering ambiguous stimuli, and that the false positives produced by the system are an important factor in accounting for religious belief. Experimental findings, however, have so far failed to lend support to current theoretical models of agency detection and have been unable to find the hypothesized perceptual biases that logically follow from them. In much the same way, theoretical models from research on sensed presence have not been corroborated by recent experimental findings, whereas earlier findings are probably best considered artifacts of flawed methodology.

With the benefit of hindsight, perhaps this is not so surprising. Most cognitive theories on agency detection were formulated at a time in which evolutionary psychology was dominated by the idea that the human mind consisted of a collection of computationally distinct and domain-specific modules (Cosmides & Tooby, 1987). At the time, the brain was considered akin to a Swiss army knife composed of a range of cognitive "programs" each specialized to handle particular types of information. Today, new developments in cognitive science encourage a stronger focus on processual and domain-general models of perception and cognition (Clark, 2016; Hohwy, 2013). As opposed

to domain-specific models, such models utilize common computational principles to process information from a range of different cognitive domains and have a strong emphasis on the developmental capability of the systems. In these models, the human brain is better thought of as a hand rather than a Swiss army knife. Although a human hand has a long evolutionary history, it is extremely adaptable and can perform a vast variety of different tasks that natural selection did not "foresee" (Heyes, 2012).

Considering the fact that agency detection research has primarily been occupied with perceptual phenomena, the lack of engagement with current proximate models of perception is striking. Without an explicit proximate perceptual model, it is difficult to make predictions about when in particular the human perceptual system should be prone to produce false positives. One of the most promising perceptual theories currently being pursued within the cognitive sciences is "predictive coding" (Friston, 2009; Friston & Kiebel, 2009; Frith, 2007). This theory states that the brain is a statistical organ that constantly tests its own hypotheses about the world through an ongoing process of error minimization. This process involves the brain attempting to predict the sensory input it receives from the world and minimize the errors that arise when inaccurate predictions are made. Predictive coding was initially intended as a theory of perception and action, but contemporary scientists and philosophers have gone further, debating whether the computational principle inherent in the theory can be extended to account for how the human mind works in general (Clark, 2016; Hohwy, 2013). Regardless of its future status as a grand theory of the mind, however, impressive amounts of empirical evidence already point in its favor as a theory of perception and action (Bubic et al., 2010; Clark, 2013, 2016; Den Ouden et al., 2012; Friston & Stephan, 2007; Hohwy, 2013, 2014; Huang & Rao, 2011). A robust model such as this may have the potential we need to solve the pending theoretical inconsistencies in agency detection research, account for the puzzling experimental findings we have encountered, and provide us with hypotheses for future experimental testing.

3. The predictive mind

In recent years, cognitive neuroscience has increasingly adopted predictive coding to model how the brain processes sensory information. Predictive coding specifies how the brain is constantly in the process of predicting incoming sensory input and thereby inferring the causes in the environment of that input; crucially, this inferential process relies heavily on prior expectations. By focusing on sensory input that does not fit its predictions – the prediction errors – the brain elegantly can come to represent the world accurately. When sensory input conflicts with predictions, prediction errors are passed up the neuronal hierarchy, allowing the brain to update its model of the world through an ongoing process of prediction error minimization (Friston, 2009; Friston & Kiebel, 2009; Frith, 2007). Predictive coding is immensely promising. Due to its reliance on a single powerful neuronal mechanism, it promises to parsimoniously explain a host of key perceptual problems and formulate them mathematically and mechanistically (Hohwy, 2013; for an introduction to the topic, I recommend Clark, 2016; Frith, 2007; Hohwy, 2013). It stands in contrast to earlier models of perception in which the brain is portrayed as a passive receiver of input that only structures and organizes the input it receives in a predominantly bottom-up manner. In contrast to this idea of a static system, what is emerging is a pro-active, hypothesis-generating machine that is constantly testing its hypotheses against the incoming signal. However, since on this kind of account, inference depends on prior learning, the brain is limited by what it already knows, which can lead to false inference and errors in some circumstances. As will be discussed below, some of these are highly relevant to the question of agency detection.

Predictive coding seeks to explain how the brain, unbeknownst to consciousness, engages in a process that approximates Bayesian inference. In Bayesian inference, a model of the world is picked or revised on the basis of how well it can predict new sensory evidence, and on the basis of how probable that model is independently of that sensory evidence. Thus, a simplified version of Bayes' rule

$P(h_1|e) \propto P(e|h_1)P(h_1)$ states that when encountering sensory evidence, the probability of a given model (*posterior probability*: $P(h_1|e)$) will be given by considering the likelihood of the sensory evidence given that the model is correct (*likelihood*: $P(e|h_1)$) and weighting this by the subjective estimate of how likely that model is *independent* of the sensory evidence *(prior probability*: $P(h_1)$). In other words, in approximating such inference, the brain can be said to assign probabilities to multiple competing models by assessing each model's weighted likelihood and prior probability. The model with the highest posterior probability is then selected, corresponding to the model that reduces prediction error the most. Importantly, the model that is ultimately picked by the brain is what is consciously experienced (Hohwy, 2013).

A few examples will serve to illustrate how this is proposed to work for perceptual cases. In the well-known rubber hand illusion, participants are asked to hide one of their hands from view, after which a rubber hand is placed at the approximate previous location of the hand they have just hidden from view. The experimenter then touches the rubber hand and the real hand in synchrony. This causes participants to experience a striking sensation of ownership over the rubber hand (Botvinick & Cohen, 1998). This illusion is nicely explained by the predictive coding framework. The illusion is produced by two models of the world coming into competition with one another. One model specifies that the participant is watching a rubber hand being touched, but that this has nothing to do with the tactile sensations felt on the hand they have hidden. The other specifies that the touch the participant feels is the result of the synchronous action they see proceeding in front of them. While the first model has the higher likelihood of being true if sensory evidence alone were considered, the second model wins, because the prior probability that visual and tactile information that co-occur in time share a common cause is very high (Botvinick & Cohen, 1998; Hohwy, 2013). In other words, the strong expectation that things that co-occur in time have a common cause drives the rubber hand illusion.

Another phenomenon that nicely illustrates how the brain "makes up its mind" is binocular rivalry. Binocular rivalry refers to the alternation of subjective perception between different stimuli simultaneously presented to each eye (Porta, 1593). For instance, a picture of a house may be presented to one of the participant's eyes, and a picture of a face to the other. Although it might be expected that participants would report seeing some kind of mishmash of a face on top of a house, what actually happens is that participants see a house for a few seconds, then a face for a few seconds, then a house, then a face, and so on. Predictive coding explains binocular rivalry by reminding us that the brain is heavily influenced by prior expectations when it tries to predict the causes of sensory input. In this case, the brain needs to explain the sensory input received through each eye. If we consider the competing models that the brain most likely has to choose between, these are (1) the mishmash model, in which the face and house occur in the same spatiotemporal position, and (2) the alternation model, in which the face and house occur in the seen space, but separated in time. While the likelihood of the mishmash model is higher than the alternation model and able to explain more of the sensory input, the prior probability of a house occupying the same space as a face is extremely low. Put simply, the brain chooses the alternation model because we never expect two objects to occupy the same space at the same time[4] (Hohwy, Roepstorff, & Friston, 2008).

Importantly, Bayesian inference takes noise and uncertainty into account.[5] This is crucial for a real world predictive mind since the world does have different contexts of uncertainty that should be reflected in inference[6] (e.g., we learn how levels of light change in different contexts and this impacts on how much we trust visual input versus prior expectations). In terms of neuronal mechanisms, such precision-weighting of prediction errors is proposed to happen through postsynaptic gain (Feldman & Friston, 2010; Friston, 2009). This means that perception will tend to be dominated by bottom-up sensory input when precision is expected to be high and by top-down prior expectations when precision is expected to be low, and that precision should vary according to context (Clark, 2016). This again has implications for when the brain is likely to commit errors and make false inferences about the state of the world. Consider, for instance, that you find yourself alone at night in a forest. Here, sensory input will be expected to have low reliability because of the

darkness, which, in turn, will facilitate top-down modulation of your experience ("the path is usually over here"). Now consider a scenario of relevance to the topic of this chapter: before leaving your home, you had heard on the radio that a psychotic serial killer was on the loose in your area, enhancing the prior probability of encountering said serial killer. Taken together, the low reliability of sensory signals combined with your prior belief would raise the likelihood that you would mistakenly perceive a rustling in the bushes for a lurking murderer.

The rubber hand illusion and binocular rivalry are both perceptual illusions that have been the subject of much debate in cognitive neuroscience, and predictive coding suggests elegant solutions to both of these perceptual phenomena (Apps & Tsakiris, 2014; Hohwy et al., 2008). At the same time, binocular rivalry and the rubber hand illusion are both illustrative cases of how the computational principles of the brain sometimes give rise to perceptual errors, much like the false positives stipulated to be produced by the human agency detection system. These phenomena highlight how reliant the human perceptual system is on expectation and how illusions can occur even in situations in which sensory information is available and apparently precise (though manipulated in other ways). When information is not reliable, the perceptual system relies to an even greater degree on top-down prior expectation, creating good conditions for false positives of different kinds. As such, the Bayesian perspective offered through predictive coding provides a promising theoretical foundation for explaining how and when humans will tend to produce false positives in agency detection.

4. Agency detection reconsidered

The research literature on agency detection generally proposes that humans have evolved a hypersensitive agency detection device that, by following the mantra "better safe than sorry," routinely raises false alarms of agents, and furthermore it has generally been asserted that this cognitive module is especially hyperactive in situations of uncertainty and when it encounters ambiguous evidence. As we saw, however, the claim that our perceptual system processes ambiguous stimuli according to the mantra "better safe than sorry" is supported neither by the experimental evidence nor by the predictive coding framework. If humans processed ambiguous information in this way, we would expect a response bias towards agent categorization as a function of noise levels in classical paradigms such as the face-house categorization task and the biological motion task, but no such tendency has yet been found (van Elk, 2015; van Elk et al., 2016). As an alternative theoretical framework, predictive coding offers an attractive explanation as to why we do not see the effects that naturally follow from HADD theory. At climbing noise levels, the face model and the house model will steadily be able to explain the same amount of sensory evidence. This means that the subjective estimate of how likely each model is *independently* of the sensory evidence (the models' prior probability) will increasingly determine which model is chosen. However, since participants in the face-house paradigm have no reason to expect more faces than houses (since they are exposed to an equal distribution of both), the face-house categorization distribution should be equally distributed. Only if participants had reason to expect more faces than houses should we expect to see a skewness in the data. For example, in a novel version of the binocular rivalry paradigm, Denison, Piazza, and Silver (2011) presented participants with identical images to both eyes before moving on to rival test image pairings. One of these test images was consistent with the preceding images while the other was not. Results revealed that participants were more likely to perceive the consistent image at the onset of binocular rivalry. That is to say, participants first saw the image they had reason to expect seeing. This suggests that the human brain is not predetermined to select agent models to account for ambiguous stimuli. Instead, the models chosen to account for ambiguous sensory evidence rely heavily on context-sensitive subjective estimates of prior probability. The same seems to be true with regard to ambiguous movements. Even when humans encounter ambiguous stimuli in the form of moving objects or entities, they do not seem to automatically follow the "better safe than sorry" mantra. For example, we do not seem to respond to any random movements as though they were agents. It is specifically

motion that seems to be driven by hidden states as opposed to resulting from external environmental causes that facilitates agency detection (Castelli, Frith, Happé, & Frith, 2002).

4.1. Agency detection in the lab

As it stands, predictive coding is a very promising framework that posits clear hypotheses for agency detection that are ready for experimental testing. Bayesian inference implies that priors are weighted more heavily when precision of sensory input is low, and in the predictive coding theory, prediction errors are weighted according to their expected precisions. This means that perception is normally dominated by bottom-up sensory input when precision is high and by top-down prior expectations when precision is low. This also means that the framework lends some truth to the claim that humans experience more false alarms of agents under conditions of ambiguity or uncertainty, as maintained by current literature on agency detection (Atran, 2002; Barrett, 2004; Guthrie, 1993). However, false alarms such as these should mostly arise in contexts in which the prior probability of agent occurrence is high. By way of example, if you were unlucky enough to find yourself in a snake pit at night and your shoe were to bump into something, you would probably experience it as a snake. If, on the other hand, you were in a stone quarry, you would probably experience a stone. In other words, predictive coding tells us that we should expect false alarms of agents in contexts in which the precision of sensory input is low (e.g., mist, darkness) and the prior probability of agent occurrence is high. CSR could benefit from moving on from the notion that humans possess a HADD, and adopting the strong alternative framework provided by predictive coding. As understood from this perspective, we are equipped not with a hyperactive agency detection device, but with a Bayesian inference machine that is vulnerable to false alarms of virtually any kind when the appropriate prior probabilities are high and the precision of sensory signals is low.

The crucial variables that need to be manipulated in order to test predictive coding against HADD-based interpretations are thus the degree of prior expectation and the reliability of sensory stimuli. While classical paradigms such as those concerned with binocular rivalry or the face-house categorization task could potentially be adapted for this task, a more attractive and ecologically valid idea would be to utilize emerging technologies such as Virtual Reality. A virtual, yet realistic world would allow for participants' experience of sensory precision to be manipulated through the virtual implementation of environmental factors such as fog, mist, sandstorms, nightfall, and so on. In such a setup, participants could be instructed to push a button as soon as they thought they detected an agent, although no virtual agents would actually be programmed into the world. Traditional HADD theory and predictive coding theory would both predict an increase in the number of agents detected in the non-reliable condition compared to the reliable condition. One could, however, introduce another level to such a paradigm by creating a second condition where participants would be asked to detect some sort of moving but non-agentive object like, say, tumbleweed or falling rocks. Again, no such objects would actually have to be present in the virtual world. In such a condition, predictive coding theory would predict an increase in tumbleweed detection in the non-reliable condition compared to the reliable condition, whereas traditional HADD theories would not. Similarly, one could also envisage a repeated trials design in which groups would be immersed in a virtual world (e.g., a forest) with either reliable or unreliable stimuli (e.g., clear or misty weather). In such a setup, some participants could be manipulated to expect a high occurrence of agents, whereas others could be manipulated to expect a low occurrence of agents. By virtue of the repeated measures design, participants would undergo the same condition multiple times, whereby a learning curve could be obtained from participants. Traditional HADD theory would hypothesize (1) that participants would report a stable amount of (false) agent detections in each trial, and (2) that reported (false) agent detections would be chronically higher in the unreliable condition compared to the reliable condition. As opposed to this, predictive coding informs us that Bayesian brains factor in both the prior expectation of a phenomenon and the reliability of the given evidence. This would mean that while we would expect all participants to move towards zero (false) agent detections as

they go through multiple trials (and learn that no agents are really present in the virtual world), participants in the high expectation condition should report a higher number of (false) agent detections compared to participants in the low expectation condition in the first trials. However, participants in the low reliability condition should learn this at a slower rate than participants in the high reliability condition, because their perception presumably will be more dominated by top-down processes in the face of unreliable stimuli over the trials.

Needless to say, such experiments would crucially depend on the ability to induce strong expectations of encountering agents in participants. One possibility would be to prime expectations through having a training phase consisting of spotting actual agents/tumbleweed under conditions of normal perceptual reliability, then simply removing these stimuli in the perceptually obscured test conditions. Such a setup would not only benefit from having a higher ecological validity than experimentally controlled lab-based studies that present stimuli on a standard computer screen; it would also allow for variations of the paradigm to systematically identify and analyze the components responsible for false positives in agency detection. One could, for instance, also choose to include a condition designed to induce fear in participants by utilizing common elements from horror games in order to test the longstanding claim that threat increases false positives in agency detection.

Another possibility would be to select participants based on their prior beliefs and see how these affect the detection of agency under conditions of sensory impoverishment, as predictive coding entails that perception is normally dominated by bottom-up sensory input when precision is high and by top-down prior expectations when precision is low. This consideration has implications for existing findings that may help point the way in developing future experimental paradigms. For instance, Persinger's God helmet findings may be entirely accounted for by considering what happens to a Bayesian brain that is put in sensory deprivation. By definition, individuals in a condition of sensory deprivation experience a general lack of sensory input. This condition is not quite as absolute as it sounds, however, since it is impossible to create complete sensory deprivation by physical means. Understood from the perspective of predictive coding, this general lack of clear and precise sensory input forces top-down predictions to impose structure on the small degree of available sensory noise (Corlett, Frith, & Fletcher, 2009). This accounts for the hallucinations that are generally observed in individuals put in sensory deprivation, where experiences of other presences are frequent (Bexton, Heron, & Scott, 1954; Corlett et al., 2009). However, such extreme conditions of uncertainty could also facilitate cognitive penetrability, such that higher-order subjectively held beliefs (such as religious beliefs) might play a greater role in determining what people actually perceive (Hohwy, 2013). Pursuing this idea, a Danish research team created a placebo God helmet paradigm, where they gave participants instructions that they would experience the sensed presence of sentient beings while being stimulated with the helmet, which was in actuality unable to induce any such experience (Andersen et al., 2014). As predicted, the team found that suggestive instructions and context combined with sensory deprivation were indeed sufficient to induce experiences of sensed presence. They also found significant differences between the three groups participating in the study, suggesting that the successful elicitation of such experiences is particularly dependent on the beliefs and background of the participants (Andersen et al., 2014). Cognitive penetrability, therefore, may well account for these group differences in which spiritists to a significantly higher degree reported experiences of sensed presence than the control group.

4.2. Supernatural agency detection

Predictive coding is compatible with a lot of the central elements in previous accounts of the human agency detection system as suggested by Guthrie, Atran, and Barrett, most saliently the proposition that people will be especially likely to have false positives of agents when they encounter ambiguous or noisy evidence. But the model puts a much stronger emphasis on expectations in the generation of false positives in agency detection, entailing a larger emphasis on the role of cultural practices in the generation of such experiences than previous models have done. If the human perceptual system is

akin to a Bayesian inference machine, then agents should not by default have a privileged place in our perceptual apparatus. Instead, we should mostly expect false alarms of agents in contexts where the precision of sensory input is low and the prior probability of agent occurrence is high.

This leaves us with the following problem: if the detection of agents does not have a privileged place in our perceptual system, then how do we explain the fact that supernatural agents are so cross-culturally prevalent? This cross-cultural supernatural agent prevalence is precisely the phenomenon that HADD theorists aimed to address when they postulated the existence of a hard-wired hyperactive agency detection mechanism. The HADD solution itself, however, entails another large problem: most religious individuals have not become religious as a result of actually *perceiving* a supernatural agent first hand (Boyer, 2001). On the contrary, the strongest predictors of religiosity are various kinds of socialization (religious upbringing, ritual attendance, religious schooling, societal levels of religiosity, and so forth). The majority of religious people have the supernatural beliefs they do, not because they have had direct perceptual experiences of supernatural agents, but because they have adopted these beliefs from other humans in their society (e.g., Gunnoe & Moore, 2002). This points to a crucial distinction between the transmission of ideas about supernatural agents on the one hand (e.g., knowing what a ghost is) and perceptual experiences of supernatural agents on the other (e.g., actually seeing a ghost). As I will describe below, it is in the dynamic relationship between conceptual transmission and perceptual experience that a revised Bayesian notion of agency detection can be meaningfully applied to account for some of the many aspects of religious belief.

The successful transmission of religious ideas is critical for the survival of any religious system. Epidemiological accounts of religious transmission hold that those religious concepts that have succeeded, often at the expense of other competing concepts, are those that were the most memorable, motivating, and convincing, thereby yielding a strong selective advantage (Boyer, 2001; Sperber, 1996). Probably the most widely cited property said to be shared by the majority of successful religious concepts is minimal counterintuitiveness (Boyer, 1994), which has been experimentally demonstrated to have mnemonic advantages (e.g., Barrett & Nyhof, 2001; Boyer & Ramble, 2001).[7] Similarly, it has been argued that certain ritual practices facilitate accurate transmission of religious ideas and concepts specifically by tampering with mnemonic cognitive processes (Schjoedt et al., 2013a, 2013b; Whitehouse, 2002). Besides being easy to remember, however, a successful religious concept must also capture the interest of potential adherents in order to facilitate and increase transmission, something that is often achieved by content rendering the concept relevant and applicable to daily human concerns. It has been argued that this is why we find such a large distribution of supernatural agents with privileged access to strategic information about human beings (Boyer, 2001; Hammer & Sørensen, 2010). Compared to supernatural agents without such privileged access, strategically informed supernatural agents will gain a selective advantage. In other words, epidemiological accounts of religious belief can explain the prevalence of supernatural agents in religious systems without invoking the human perceptual system at all, rendering the hypothesis that supernatural agent beliefs are evidence for a hard-wired agency detection device within the perceptual system superfluous.

However, in a similar manner, it seems safe to assume that religious systems that can facilitate actual experiences that serve to confirm the existence of relevant agents to believers should, all else being equal, gain a selective advantage over systems that do not enjoy such possibilities. This is where false positives in agency detection become particularly relevant to the scientific study of religion. For instance, ghost sightings in locations reputed to be haunted (Bader, Mencken, & Baker, 2011), spirit possession (a cross-culturally recurrent phenomenon) (Cohen & Barrett, 2008; Lewis, 2003), and direct experiences of the Holy Ghost in Pentecostal churches (Luhrmann, Nusbaum, & Thisted, 2010) all serve to maintain and strengthen connected supernatural representations, both via the direct experiences themselves and via the testimonies produced by "witnesses."

But how can a religious system facilitate these experiences? While it may initially seem trivial to claim that religious and spiritual systems attempt to alter expectations of where one might encounter

ancestors, angels, gods, and ghosts (we all know that whether it be haunted houses, burial grounds, monasteries, or temples, religious individuals are primed to expect apparitions in some places more than others), this is in fact another way of saying that religious systems communicate estimates of the distribution of supernatural agents in certain environments or, rephrased in Bayesian terms, *religious systems often provide their users with estimates of the prior probability of the occurrence of supernatural agents*. However, while other means exist, the bulk of such expectations are often induced through exposure to higher-order verbal or written information such as sermons or scripture. This begs the crucial question: can higher-order information of this sort really affect lower-level perceptual processes?

Experimental evidence suggests that it can (Lupyan, 2015). In one recent study, it was found that exposure to relevant language led people to perceive otherwise invisible stimuli (Lupyan & Ward, 2013). Lupyan and Ward used a continuous flash-suppression (CFS) paradigm, a variant of binocular rivalry, in which participants are shown images that are suppressed in a way that renders the pictures impossible to see. Interestingly, when participants were aurally exposed to a lexical item designating the suppressed image, this was enough to make the image visible to the participants. Similarly, higher-order information has also recently been shown to directly affect the perceived *reliability* of stimuli (Schiffer, Siletti, Waszak, & Yeung, 2016). In a novel probabilistic reinforcement learning task, Schiffer and colleagues provided participants with *instructions* about the reliability of feedback or volatility of the environment, while at the same time manipulating the *actual* reliability of feedback and volatility of the environment. Their results revealed that participants adapted their behavior faster to changes in the environment when they received instructions indicating that negative feedback was particularly informative. Conversely, in instances where the environment did not in fact change, explicit instruction had a negative effect on adaptability. In other words, instructions led participants to perform more poorly if they were incongruent with the task setup and better if they were congruent. In sum, experiments such as these suggest that higher-order information in the form of verbal instructions does in fact have effects on low-level perceptual processes and estimates of sensory reliability (for more studies on the effects of high-level information on lower-level processes, see Lupyan, 2015).

These insights provide one of several answers to the apparent paradox of how humans can gain priors about supernatural agents in the first place (Schjoedt & Andersen, 2016). Religious teachings and narratives seem to constitute one source of priors, resembling the phenomenon of "top-top processing" (Roepstorff & Frith, 2004) or "supra-top-down processing" (Shea et al., 2014); in other words, they provide *script-sharing for perception (and action)* between individuals. At other times, priors seem to be gradually induced through online guided processing of ambiguous stimuli (Schjoedt & Andersen, 2016). Take, for instance, recent studies of Christian evangelical prayer practice in which believers come to infer the presence of a supernatural agent through guided and repeated attempts to interpret bodily states (Luhrmann et al., 2010). In such scenarios, religious experts seem to guide the perceptions of believers in order to slowly transform their priors to match ambiguous bodily sensory states. Importantly, when prior beliefs match sensory input ("the Holy Spirit feels like the tingly feeling I'm feeling now"), predictive coding suggests that these prior beliefs will grow stronger. This repeated guided conjunction of ambiguous sensation with supernatural interpretation may explain why such religious practices can lead to a point where believers eventually come to experience the presence of a supernatural agent without the need for online guidance at all (Schjoedt & Andersen, 2016).

Alongside the higher-order manipulation of expectations, it is also crucial to appreciate that a wide range of religious practices actively encourage members to engage in different forms of sensory deprivation. Indeed, there are striking similarities in some of the most basic and widespread religious practices, namely prayer and meditation, in that religious individuals are universally encouraged to close their eyes and seek quiet surroundings. In some cases, such encouragements extend to even more extensive forms of deprivation such as nightly rituals, social seclusion, voluntary inhibition of motor actions, and so on (La Barre, 1972). From the perspective of predictive coding, such

practices force the brain to rely more on prior expectations than it otherwise would. When combined with a critical amount of culturally induced expectation, this may just be enough to create false positives in agency detection.

Taken together, this leads to the conclusion that many religious systems actively manipulate the relationship between expectations and sensory information to target the blind angles of the perceptual system in ways that may facilitate false positives in agency detection. Religious systems seem to do this by inducing expectations of supernatural agents in their users through the transmission of religious ideas, while frequently also encouraging periods where those very same users' sensory systems are inhibited. This produces a cyclical feedback loop where the transmission of religious ideas facilitates culturally endorsed supernatural agent experiences, further strengthening expectations that such agents exist and may be encountered, which in turn strengthens the religious system itself. This somewhat corresponds to the idea put forward by previous HADD researchers that sensed encounters with supernatural agents may well encourage, strengthen, and spread pre-existing beliefs in supernatural agents (Barrett & Lanman, 2008), with the important addition that pre-existing beliefs can now additionally, and perhaps more importantly, be viewed as independent drivers of those very same experiences.

A possible objection that could be raised at this point is that it seems ironic to suggest that false positives in agency detection could arise in a system that is fundamentally designed to detect and correct errors. Why, one might even ask, are such errors not immediately identified and expunged? The answer is simply that if there is no reliable sensory feedback, the perceptual system will not flag an error. Generally speaking, and in the course of ordinary activity, if sensory reliability is low, humans will actively investigate the environment in an effort to reduce the unreliability of incoming stimuli (Friston, Mattout, & Kilner, 2011). For instance, if we were to find ourselves lying awake at night imagining that an ominous tapping on our window was a burglar attempting to force it open, further investigation might quickly reveal that the actual culprit was a tree branch caught in the wind. Consequently, false positives in agency detection should mostly transform into stable and lasting experiences of supernatural agency in situations where the individual is either unmotivated to explore the environment, unable to or inhibited from exploring the environment, or when the perceived agent is expected to only be perceptible for a short period of time. This may explain why most religious belief systems contain supernatural agents that are predominantly imperceptible, while those that are perceptible tend to be portrayed as elusive and evasive. Furthermore, the well-attested flexibility and empirical unfalsifiability typical of supernatural agent concepts may well mean that they have succeeded precisely through developing in a manner that makes them impervious to Bayesian tendencies to revise error. If we combine this with the fact that widespread practices such as meditation and prayer predominantly inhibit active inference from the surrounding environment, it seems likely that certain religious practices and supernatural agent beliefs may have co-evolved in a way that targets the Bayesian "sweet spot" of sensory ambiguity. We seldom hear, for instance, of ghosts that show themselves in broad daylight; instead, they almost invariably roam the earth at night or in dimly lit places. In other words, it seems that successful concepts of perceptible supernatural agents, beyond simply being memorable and motivating, may also have evolved in a way that targets the blind angle of the human perceptual system, reinforcing their believability and transmissibility by rendering actual sensory experiences possible under the right circumstances.

Similarly, some technologies seem to have evolved explicitly to facilitate experiences of supernatural presence. For example, the success of spiritualism in America and Europe during the second half of the nineteenth century can be largely attributed to the invention of several new spiritual communication devices, including writing planchettes, dial plates, and talking boards (Lehmann, 1920). Such technologies typically leave practitioners struck by the sensation that they are not responsible for the movement of the devices, in turn making the idea that a supernatural agent is responsible an attractive option (Wegner, 2002). In cognitive science, this experience of controlling one's actions is normally referred to as the "sense of agency." A range of experiments, typically interpreted within a predictive coding framework, suggest that the feeling of control is computed by "predicting the

consequences of current actions, and comparing these predictions to actual outcomes" (Haggard & Chambon, 2012, p. 390). If sensory feedback matches predictions, people experience control over the action. If not, people experience a decrease in their sense of agency (Frith, 2005). Consistent with these findings, a recent field experiment among Ouija board enthusiasts employing mobile eye-tracking technology found that Ouija users were significantly less able to predict the movements of the planchette during attempts to communicate with supernatural entities compared to the control condition, where participants were asked to deliberately spell out words with the planchette themselves (Andersen et al., 2017). This is yet another example of how cultural practices and technologies can manipulate the sensory system and create powerful and persuasive experiences of supernatural agents.

Beyond this, predictive coding, as a domain-general model of perception, entails that such culturally instantiated manipulation of the sensory system is not confined to experiences of supernatural agents. The computational principles of predictive coding extend to a range of other religious and spiritual practices not necessarily concerned with the detection of agents. In Reiki healing, for example, practitioners will often perform so-called "palm-healing" where the practitioner places her hands slightly above the patient's body, claiming to transfer "universal energy." The patients receiving such treatment often report that they can feel the energy being transferred although physical contact occurs at no point during the session. This interesting perceptual phenomenon obviously adds to the credibility of the healing session, and we should also note that the laying on of hands is not an uncommon phenomenon in other religious and spiritual traditions. Seen through the lens of predictive coding, such perceptions arise through the strong expectation of receiving tactile stimulation when another person places his or her hands on or even simply over you. The strong expectations of the patient drive the experience of tingling flowing from the Reiki practitioners' hands. As illustrated, employing predictive coding as a theoretical framework ultimately means that false positives in agency detection can be contained within the same model as false positives of virtually any other perceptual phenomenon. This means that predictive coding is not only a promising model to employ for research on agency detection, but also has great potential for explaining a range of other perceptual experiences relevant to the cognitive science of religion.

5. Conclusion

Predictive coding is currently one of the most promising models of human perception and action. It has powerful theoretical arguments backing it, mounting empirical evidence in support of it, unifying power, and great potential in terms of explanatory scope. In this article, I have argued that predictive coding offers an attractive unifying framework that solves the theoretical inconsistencies and puzzling experimental findings that have so far bedeviled research on agency detection. As an alternative to a hypersensitive cognitive module, predictive coding presents a general Bayesian inference machine that is vulnerable to false alarms of virtually any phenomenon when prior probability is high and the precision of sensory signals is low. This means that false positives in agency detection can be contained in the same model as false positives in a range of other perceptual phenomena relevant to the cognitive science of religion. Predictive coding entails a stronger emphasis on expectations in the generation of such perceptions, and it entails an encouragement to CSR to focus more on the role of cultural practices and technologies in the generation of perceptual experiences deemed religious or spiritual.

Notes

1. Importantly, the concept of "agency detection" refers to the *perceptual* process of detecting agents, and is not to be confused with higher-order cognitive processes such as "anthropomorphism" (the attribution of human-like characteristics, motivations, intentions, and emotions to the real or imagined behavior of non-human agents

[Epley, Waytz, & Cacioppo, 2007]) or "mentalizing" (the cognitive process of reading the content of other minds [Frith & Frith, 2003]).

2. Guthrie argues that human perceptual representations depend on three factors: (1) the phenomenon interpreted must somehow correspond to the representation evoked; (2) the representation must reflect the likelihood of occurrence of the phenomenon; and (3) the representation must be of importance to the observer. He goes on to say that because agents are of the utmost importance to humans, and because human agents in particular can be the cause of such a wide variety of phenomena, a vast host of phenomena are therefore likely to be caused by humans, leading in turn to the production of false positives in numerous contexts (Guthrie, 1980, 1993).
3. In a letter to the editor, Persinger and Koren later responded that the alleged replication by Granqvist and colleagues was not a true replication. In their response, Persinger and Koren argue that Granqvist and colleagues used the wrong equipment, did not expose the participants long enough for magnetic stimulation to occur, and that the room size used by the Swedish research team was too small (Persinger & Koren, 2005). Granqvist and colleagues replied that they had used the equipment from Persinger's own lab and that Persinger himself had personally ensured the correct calibration of the equipment. Furthermore, the team replied that Persinger had personally recommended the administered duration of the magnetic stimulus and that the room size used for the experiment in fact very closely resembled that of Persinger's own lab, although the dimensions of the room were erroneously described in their original article. The Swedish research team provided e-mail correspondence as proof of these claims (Larsson, Larhammar, Fredrikson, & Granqvist, 2005).
4. This explanation appeals not just to aspects of simple Bayesian inference, but more specifically to the predictive coding element. It proposes that the alternation between the two images is a result of prediction error dynamics, where unexplained prediction error from the suppressed stimulus eventually demands to be explained and thereby forces the perceptual alternation. Further, this explanation has a hierarchical aspect, according to which a higher-level expectation of the rate of change of the inferred causes controls the dynamics of the lower-level sensory attributes. Overall, this can lead to adaptation or weakening of the prior expectation over time, which helps explain why the prediction error from the competing, currently unperceived stimulus can begin to dominate inference.
5. This is the property of the denominator in Bayes' rule (the marginal likelihood), which is not alluded to in the simplified version of Bayes' rule mentioned above. For a comprehensible chapter on Bayesian models of cognition, see Griffiths, Kemp, and Tenenbaum (2008).
6. In some schemes, which suggest process theories to implement Bayesian inference, this aspect of inference is dealt with separately and given a pivotal role (Friston, 2010). The reasoning here is that the noise can be conceived as the variance of a probability distribution, the inverse of which is the precision, and thereby (normal) distributions can be represented through their sufficient statistics which for normal distributions becomes their means and precisions. Now one can think of precisions on their own, and consider prediction error minimization for them, which means that precisions can be learnt in a process that approximates Bayesian inference and facilitate reliable inference in a world with varying levels of uncertainty. In a hierarchical setting, this means that inference becomes context sensitive; it can vary the relative weighting of priors and likelihood according to the expected precision of the input.
7. Importantly, however, it has recently been argued that it is non-trivial to infer from existing experimental evidence that the cultural ubiquity of religious concepts can be explained by minimal counterintuitiveness theory. For a critical review of research on minimally counterintuitive concepts, see Purzycki and Willard (2016).

Acknowledgments

The author gratefully acknowledges the comments from the editor and two anonymous reviewers, which benefited the manuscript greatly. The author is also grateful for the helpful and insightful comments on the manuscript provided by Jesper Sørensen, Religion, Cognition & Culture, Aarhus University, Denmark; Andreas Roepstorff, Interacting Minds Centre, Aarhus University, Denmark; Joshua Skewes, Interacting Minds Centre, Aarhus University, Denmark; Hugh Turpin, Institute of Cognition and Culture, Queens University Belfast, Ireland; and Jacob Hohwy, Cognition & Philosophy Lab, Monash University, Australia.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by Aarhus Universitet, Interacting Minds Centre [grant number 10544].

References

Andersen, M., Nielbo, K. L., Schjoedt, U., Pfeiffer, T., Roepstorff, A., & Sørensen, J. (2017). Predictive minds in Ouija boards sessions. Manuscript submitted for publication.

Andersen, M., Schjoedt, U., Nielbo, K. L., & Sørensen, J. (2014). Mystical experience in the lab. *Method and Theory in the Study of Religion*, *26*, 217–245.

Apps, M. A. J., & Tsakiris, M. (2014). The free-energy self: A predictive coding account of self-recognition. *Neuroscience & Biobehavioral Reviews*, *41*, 85–97.

Atran, S., & Norenzayan, A. (2004). Religions evolutionary landscape: Counterintuition, commitment, compassion, communion. *Behavioural and Brain Sciences*, *27*(6), 713–770.

Atran, S. (2002). *In gods we trust: The evolutionary landscape of religion*. New York, NY: Oxford University Press.

Bader, C., Mencken, F. C., & Baker, J. O. (2011). *Paranormal America: Ghost encounters, UFO sightings, bigfoot hunts, and other curiosities in religion and culture*. New York: NYU Press.

Barnes, K., & Gibson, N. J. S. (2013). Supernatural agency: Individual difference predictors and situational correlates. *International Journal for the Psychology of Religion*, *23*, 42–62.

Barrett, J. (1999). Theological correctness: Cognitive constraint and the study of religion. *Method & Theory in the Study of Religion*, *11*, 325–339.

Barrett, J. (2000). Exploring the natural foundations of religion. *Trends in Cognitive Sciences*, *4*(1), 29–34.

Barrett, J. (2004). *Why would anyone believe in god?* Walnut Creek, CA: Altamira Press.

Barrett, J. (2007). Cognitive science of religion: What is it and why is it? *Religion Compass*, *1*(6), 768–786.

Barrett, J. (2011). Cognitive science of religion: Looking back, looking forward. *Journal for the Scientific Study of Religion*, *50*(2), 229–239.

Barrett, J. L., & Keil, F. C. (1996). Conceptualizing a nonnatural entity: Anthropomorphism in god concepts. *Cognitive Psychology*, *31*, 219–247.

Barrett, J. L., & Lanman, J. A. (2008). The science of religious beliefs. *Religion*, *38*, 109–124.

Barrett, J. L., & Nyhof, M. A. (2001). Spreading non-natural concepts: The role of intuitive conceptual structures in memory and transmission of cultural materials. *Journal of Cognition and Culture*, *1*(1), 69–100.

Bexton, W. H., Heron, W., & Scott, R. H. (1954). Effects of decreased variation in the sensory environment.. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *8*, 70–76.

Bloom, P., & Veres, C. (1999). The perceived intentionality of groups. *Cognition*, *71*, B1–B9.

Botvinick, M., & Cohen, J. (1998). Rubber hand "feel" touch that eyes see. *Nature*, *391*(6669), 756.

Boyer, P. (1994). *The naturalness of religious ideas: A cognitive theory of religion.* Los Angeles: University of California Press.

Boyer, P. (1996). Cognitive limits to conceptual relativity: The limiting-case of religious categories. In J. Gumperz, & S. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 203–231). Cambridge, UK: Cambridge University Press.

Boyer, P. (2001). *Religion explained - The human instincts that fashion gods, spirits and ancestors.* London: Vintage.

Boyer, P., & Ramble, C. (2001). Cognitive templates for religious concepts: Cross-cultural evidence for recall of counter-intuitive representations. *Cognitive Science*, *25*(4), 535–564.

Bubic, A., Von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4*, 25.

Bulbulia, J., & Schjoedt, U. (2013). The neural basis of religion. In F. Krueger (Ed.), *The neural basis of human belief systems* (pp. 169–190). Hove: Psychology Press.

Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, *125*, 1839–1849.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 1–73.

Clark, A. (2016). *Surfing uncertainty: Prediction, action and the embodied mind.* New York, NY: Oxford University Press.

Cohen, E., & Barrett, J. L. (2008). Conceptualizing spirit possession: Ethnographic and experimental evidence. *Ethos*, *36*(2), 246–267.

Corlett, P. R., Frith, C. D., & Fletcher, P. C. (2009). From drugs to deprivation: A Bayesian framework for understanding models of psychosis. *Psychopharmacology*, *206*(4), 515–530.

Cosmides, L., & Tooby, J. (1987). From evolution to behavior: Evolutionary psychology as the missing link. In J. Dupré (Ed.), *The latest on the best: Essays on evolution and optimality* (pp. 276–306). Cambridge, MA: MIT Press.

Csibra, G., Gergely, G., Bíró, S., Koós, O., & Brockbank, M. (1999). Goal attribution without agency cues: The perception of "pure reason" in infancy. *Cognition*, *72*, 237–267.

Denison, R. N., Piazza, E. A., & Silver, M. A. (2011). Predictive context influences perception selection during binocular rivalry. *Frontiers in Human Neuroscience*, *5*, 166.

Den Ouden, H. E., Kok, P., & De Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, *3*, 548.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886.
Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*(215), 1–23.
Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.
Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221.
Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*(1-2), 137–160.
Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*(3), 417–458.
Frith, C. (2005). The self in action: Lessons from delusions of control. *Consciousness and Cognition*, *14*(4), 752–770.
Frith, C. (2007). *Making up the mind: How the brain creates our mental world*. Oxford: Blackwell Publishing.
Frith, U., & Frith, C. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*, 459–473.
Gervais, W., & Henrich, J. (2010). The Zeus problem: Why representational content biases cannot explain faith in gods. *Journal of Cognition and Culture*, *10*, 383–389.
Gervais, W. M., Willard, A. K., Norenzayan, A., & Henrich, J. (2011). The cultural transmission of faith: Why innate intuitions are necessary, but insufficient, to explain religious belief. *Religion*, *41*(3), 389–410.
Granqvist, P., Fredrikson, M., Unge, P., Hagenfeldt, A., Valind, S., Larhammar, D., & Larsson, M. (2005). Sensed presence and mystical experiences are predicted by suggestibility, not by the application of transcranial weak complex magnetic fields. *Neuroscience Letters*, *379*, 1–6.
Granqvist, P., & Larsson, M. (2006). Contribution of religiousness in the prediction and interpretation of mystical experiences in a sensory deprivation context: Activation of religious schemas. *The Journal of Psychology*, *140*(4), 319–327.
Green, A. (2015). The mindreading debate and the cognitive science of religion. *Sophia*, *54*, 61–75.
Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 59–100). New York: Cambridge University Press.
Gunnoe, M. L., & Moore, K. A. (2002). Predictors of religiosity among youth aged 17-22: A longitudinal study of the national survey of children. *Journal for the Scientific Study of Religion*, *41*(1), 613–622.
Guthrie, S. (1993). *Faces in the clouds: A new theory of religion*. New York: Oxford University Press.
Guthrie, S. (1996). Religion: What is it? *Journal for the Scientific Study of Religion*, *35*(4), 412–419.
Guthrie, S. (2002). Animal animism: Evolutionary roots of religious cognition. In I. Pyysiäinen & V. Anttonen (Eds.), *Current approaches in the cognitive science of religion* (pp. 38–67). London: Continuum.
Guthrie, S. (2013). Spiritual beings: A Darwinian, cognitive account. In *The handbook of contemporary animism* (pp. 353–358). Durham: Acumen Publishing.
Guthrie, S. (forthcoming). Anthropology and anthropomorphism in religion. In H. Whitehouse & J. Laidlaw (Eds.), *The salvaged mind: Social anthropology, religion and the cognitive sciences*. Durham, NC: Carolina Academic Press.
Guthrie, S. (1980). A cognitive theory of religion. *Current Anthropology*, *21*(2), 181–203.
Haggard, P., & Chambon, V. (2012). Sense of agency. *Current Biology*, *22*(10), R390–R392.
Hammer, O., & Sørensen, J. (2010). *Religion - i psyke og samfund*. Aarhus: Aarhus University Press.
Heider, F., & Simmel, S. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*, 243–259.
Hermans, C. A. M. (2015). Towards a theory of spiritual and religious experiences: A building block approach of the unexpected possible. *Archive for the Psychology of Religion*, *37*, 141–167.
Heyes, C. (2012). New thinking: The evolution of human cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 2091–2096.
Hill, D. R., & Persinger, M. A. (2003). Application of transcerebral, weak (1 microT) complex magnetic fields and mystical experiences: Are they generated by field-induced dimethyltryptamine release from the pineal organ? *Perceptual and Motor Skills*, *97*, 1049–1050.
Hohwy, J. (2007). The sense of self in the phenomenology of agency and perception. *Psyche*, *13*(1), 1–20.
Hohwy, J. (2013). *The predictive mind*. Croydon: Oxford University Press.
Hohwy, J. (2014). The self-evidencing brain. *Noûs*, 1–27.
Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*, 687–701.
Huang, Y., & Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580–593.
Hufford, D. J. (1982). *The terror that comes in the night: An experience-centered study of supernatural assault traditions*. Philadelphia: University of Pensylvania Press.
La Barre, W. (1972). Hallucinations and the shamanantic origins of religion. In P. T. Furst (Ed.), *The flesh of the gods* (pp. 261–278). New York, NY: Praeger.

Larsson, M., Larhammar, D., Fredrikson, M., & Granqvist, P. (2005). Reply to M.A. Persinger and S. A. Koren's response to Granqvist et al. "Sensed presence and mystical experiences are predicted by suggestibility, not by the application of transcranial weak magnetic fields". *Neuroscience Letters*, *380*(3), 348–350.
Lawson, T. E., & McCauley, R. N. (1990). *Rethinking religion: Connecting cognition and culture*. Gateshead: Cambridge University Press.
Lehmann, A. (1920). *Overtro og trolddom*. Copenhagen: Frimodts forlag.
Lewis, I. M. (2003). *Ecstatic religion: A study of shamanism and spirit possession*. London: Routledge.
Luhrmann, T. M., Nusbaum, H., & Thisted, R. (2010). The absorption hypothesis: Learning to hear god in evangelical Christianity. *American Anthropologist*, *112*(1), 66–78.
Lupyan, G. (2015). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology*, *6*(4), 547–569.
Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences*, *110*(35), 14196–14201.
Nielbo, K. L., & Sørensen, J. (2013). Prediction error during functional and non-functional action sequences: A computational exploration of ritual and ritualized event processing. *Journal of Cognition and Culture*, *13*(3-4), 347–365.
Persinger, M., Bureau, Y. R. J., Peredery, O. P., & Richards, P. M. (1994). The sensed presence as right hemispheric intrusions into the left hemispheric awareness of self: An illustrative case study. *Perceptual and Motor Skills*, *78*, 999–1009.
Persinger, M., & Koren, S. A. (2005). A response to Granqvist et al. "Sensed presence and mystical experiences are predicted by suggestibility, not by the application of transcranial weak magnetic fields". *Neuroscience Letters*, *380*(3), 346–347.
Persinger, M. (1983). Religious and mystical experiences as artifacts of temporal lobe function: A general hypothesis. *Perceptual and Motor Skills*, *57*, 1255–1262.
Persinger, M. (2001). The neuropsychiatry of paranormal experiences. *Neuropsychiatric Practice and Opinion*, *13*(4), 515–524.
Persinger, M. (2002). Experimental simulation of the god experience: Implications for religious beliefs and the future of the human species. In R. Joseph (Ed.), *Neurotheology: Brain, science, spirituality, religious experience* (pp. 279–292). San Jose, CA: University Press.
Persinger, M., Saroka, K. S., Koren, S. A., & St-Pierre, L. S. (2010). The electromagnetic induction of mystical and altered states within the laboratory. *Journal of Consciousness Exploration & Research*, *1*(7), 808–830.
Porta, J. B. (1593). *De Refractione. Optices Parte. Libri Novem*. Naples: Salviani.
Premack, D., & Premack, A. (1995). Origins of social competence. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 205–218). Cambridge, MA: MIT Press.
Purzycki, B. G., & Willard, A. K. (2016). MCI theory: A critical discussion. *Religion, Brain & Behavior*, *6*(3), 207–248.
Riekki, T., Lindeman, M., Aleneff, M., Halme, A., & Nuortimo, A. (2013). Paranormal and religious believers are more prone to illusory face perception than skeptics and Non-believers. *Applied Cognitive Psychology*, *27*, 150–155.
Rochat, P., Morgan, R., & Carpenter, M. (1997). Young infants' sensitivity to movement information specifying social causality. *Cognitive Development*, *12*, 537–561.
Roepstorff, A., & Frith, C. (2004). What's at the top in the top-down control of action? Script-sharing and "top-top" control of action in cognitive experiments. *Psychological Research*, *68*(2-3), 189–198.
Ruohonen, J. (1998). *Transcranial magnetic stimulation: Modelling and new techniques* (Doctoral dissertation). Department of Engineering Physics and Mathematics, Helsinki University of Technology.
Schiffer, A. M., Siletti, K., Waszak, F., & Yeung, N. (2016). Adaptive behaviour and feedback processing integrate experience and instruction in reinforcement learning. *NeuroImage*, *146*, 626–641. doi:10.1016/j.neuroimage.2016.08.057
Schjoedt, U., & Andersen, M. (2016). How does religious experience work in predictive minds? *Religion, Brain & Behavior*. Authors copy ahead of print.
Schjoedt, U., Sørensen, J., Nielbo, K. L., Xygalatas, D., Mitkidis, P., & Bulbulia, J. (2013a). Cognitive resource depletion in religious interactions (target article). *Religion, Brain & Behavior*, *3*(1), 39–55.
Schjoedt, U., Sørensen, J., Nielbo, K. L., Xygalatas, D., Mitkidis, P., & Bulbulia, J. (2013b). The resource model and the principle of predictive coding: A framework for analysing proximate effects of ritual. *Religion, Brain & Behavior*, *3*(1), 79–86.
Schjoedt, U. (2009). The religious brain: A general introduction to the experimental neuroscience of religion. *Method and Theory in the Study of Religion*, *21*, 310–339.
Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193.
Solomonova, E., Frantova, E., & Nielsen, T. (2011). Felt presence: The uncanny encounters with the numinous other. *AI & Society*, *26*(2), 171–178.
Sperber, D. (1975). *Rethinking symbolism* (No. 11). Gateshead: CUP Archive.
Sperber, D. (1996). *Explaining culture: A naturalistic approach*. Oxford: Blackwell.

Taves, A. (2009). *Religious experience reconsidered: A building-block approach to the study of religion and other special things*. Princeton, NJ: Princeton University Press.
Taves, A., & Asprem, E. (2017). Experience as event: Event cognition and the study of (religious) experiences. *Religion, Brain & Behavior*, *7*, 43–62.
van Elk, M. (2013). Paranormal believers are more prone to illusory agency detection than skeptics. *Consciousness and Cognition*, *22*, 1041–1046.
van Elk, M. (2015). Perceptual biases in relation to paranormal and conspiracy beliefs. *Plos One*, *10*(6), e0130422.
van Elk, M., Rutjens, B. T., van der Pligt, J., & van Harrevel, F. (2016). Priming of supernatural agent concepts and agency detection. *Religion, Brain & Behavior*, *6*(1), 4–33.
Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge: MIT Press.
Whitehouse, H. (1992). Memorable religions: Transmission, codification, and change in divergent Melanesian contexts. *Man*, *27*(4), 777–797.
Whitehouse, H. (2002). Modes of religiosity: Towards a cognitive explanation of the sociopolitical dynamics of religion. *Method and Theory in the Study of Religion*, *14*(3/4), 293–315.

COMMENTARIES

Predictive processing and the problem of (massive) modularity

Egil Asprem

Department of Ethnology, History of Religions, and Gender Studies, Stockholm University, Stockholm, Sweden

A theoretical shift is underway in the cognitive science of religion. As an increasing number of researchers are turning to the neurocognitive theory of predictive processing (e.g., Andersen, Schjoedt, Nielbo, & Sørensen, 2014; Asprem, 2017; van Elk & Aleman, 2017; Hermans, 2015; Nielbo & Sørensen, 2013; Schjoedt et al., 2013; Schjoedt & Andersen, 2017; Taves & Asprem, 2017), aspects of the CSR "standard model" are coming under pressure. Andersen's target article is an excellent example of this trend. Its critique of agency detection illustrates what stands to be gained by making this theoretical shift, but it also brings to light a budding conflict with the discipline's foundations in evolutionary psychology. As Andersen demonstrates, the predictive processing account offers a coherent explanatory picture, a framework for generating and testing hypotheses experimentally, and appears to solve remaining inconsistencies in empirical work on HADD. It also points out a clear role for humanist scholars, by highlighting the place of "culture" and prior learning in perception. However, this focus on learning also opens up a discussion about the role of innate, functionally specialized modules selected for in humanity's ancestral environments.

The success of conceptual shifts is judged as much by what is rejected as what is introduced. Conceptual shifts tend to involve "Kuhn loss" – when an older paradigm successfully explained some phenomenon which the new paradigm does not (Kuhn, 1970, pp. 99–100). Sometimes, new paradigms have to look backward and reconsider older explanations in order to make progress (see Midwinter & Janssen, 2012). While I am enthusiastic about the theoretical reorientation suggested by predictive processing, it is important not to throw out the baby with the bathwater. The bathwater may have become stale and murky by the expanding number of specialized modules, but in it swims a precious baby that must be saved: the idea that minds are created and shaped by natural selection.

On first glance, two aspects of Andersen's account of agency detection challenge these foundations: (1) the role of domain-general over domain-specific processing; and (2) the role of top-down effects based on prior experience and learning. The first appears to conflict with *functional specialization*, while the second questions the *innateness* of modular processing.

Regarding the first, there is a crucial distinction to be made between the Bayesian algorithm, its implementation(s) in the brain, and the overall architecture of the cognitive system (cf. Clark, 2013,

CONTACT Egil Asprem egil.asprem@rel.su.se

p. 194). Consider old-fashioned Fodorian modularity (Fodor, 1983): we might agree that the sensory modules at the cognitive periphery enjoy some degree of domain specificity (in the sense of processing different kinds of stimuli), while holding that each modality processes those stimuli in a probabilistic, Bayesian way. Although the *encapsulation* of these sensory modalities has come into question due to cross-modal and cognitive penetrability effects – nicely accounted for by predictive processing (Clark, 2016; Hohwy, 2013) – there is clearly room for some degree of functional specialization, both in the neural implementation and the wider architecture of the predictive mind.

What then about innateness? The growing evidence of cognitive penetrability and top-down learning effects stressed by Andersen's account of agent detection clearly complicates the picture of robustly innate modules – and the CSR constructs based on them (e.g., HADD, MCI concepts). However, one should note that what massive modularity theorists propose are innate *learning systems based on inbuilt biases* (Tooby & Cosmides, 1992). Things may therefore not be as black and white as Andersen presents it when suggesting that we should ditch the very idea of HADD because the domain-general predictive algorithm does not provide a privileged place for agents. This all hinges on how prior probabilities are implemented in the cognitive system: if they result from experience alone (i.e., an organism's past attempts at predicting its environment), then the mind would, indeed, be an unbiased blank slate as far as agents are concerned. But nothing in the predictive processing story forces us to adopt this particular view. In fact, there are good reasons to think that some priors should be innate and evolved rather than learned by each individual organism. The so-called "dark room" problem exemplifies this: if cognition is all about minimizing surprises (prediction error), how come we don't tend to isolate ourselves in dark, quiet caves and let the brain predict itself to near perfection? The short answer is that the brain does not expect a complete absence of sensory stimuli: that is why sensory deprivation may induce hallucinatory experiences (e.g., Corlett, Frith, & Fletcher, 2009), an effect well known to religious practices (see, e.g., Ustinova, 2009). So where do these basic expectations come from? From natural selection: if we didn't expect certain kinds of stimuli to begin with, we would crawl into dark caves and die (cf. Friston, 2013, p. 213).

Andy Clark also opens the door for innate priors of various kinds: in the shape of "hyperpriors" ("priors upon priors," organizing our experience in a Kant-like manner; Clark, 2016, p. 175) and "embodied biases" (evolved physiological structures functioning as inbuilt models of the world; ibid., 176). If we start thinking about priors not simply as what has been learned in previous iterations of predictive sensory-motor engagement with the world, but also as implemented by evolved physiology, it is (still) not inherently implausible that (say) an agent hypothesis for certain types of stimuli might come with a higher prior probability by natural design.

As long as such innate priors can be tweaked, shaped, strengthened, or weakened by cultural variation, however, their existence makes little difference to Andersen's methodological argument. But in the pursuit of a unified theoretical framework for CSR, we should seek to reconceptualize modularity in predictive processing terms rather than throw it out altogether. One possible approach is illustrated by van Elk and Aleman (2017), who seek a unified, predictive processing view of how several specific brain mechanisms associated with specific tasks (e.g., the "theory-of-mind network" and "mentalizing about gods") collectively account for phenomena deemed religious/spiritual. Approaches of this type hold great promise in connecting neurocognitive proximate mechanisms with evolutionary distal causes.

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

Egil Asprem http://orcid.org/0000-0001-9944-1241

References

Andersen, M., Schjoedt, U., Nielbo, K. L., & Sørensen, J. (2014). Mystical experience in the lab. *Method and Theory in the Study of Religion*, *26*, 217–245.
Asprem, E. (2017). Explaining the esoteric imagination: Towards a theory of kataphatic practice. *Aries*, *17*(1), 17–50.
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.
Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.
Corlett, P. R., Frith, C. D., & Fletcher, P. C. (2009). From drugs to deprivation: A Bayesian framework for understanding models of psychosis. *Psychopharmacology*, *206*(4), 515–530.
Fodor, J. (1983). *The modularity of mind*. Cambridge: MIT Press.
Friston, K. (2013). Active inference and free energy. *Behavioral and Brain Sciences*, *36*(3), 212–213.
Hermans, C. A. M. (2015). Towards a theory of spiritual and religious experiences: A building block approach of the unexpected possible. *Archive for the Psychology of Religion*, *37*, 141–167.
Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.
Kuhn, T. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.
Midwinter, C., & Janssen, M. (2012). Kuhn losses regained: Van Vleck from spectra to susceptibilities. arXiv: 1205.0179v1 [physics.hist-ph].
Nielbo, K. L., & Sørensen, J. (2013). Prediction error during functional and non-functional action sequences: A computational exploration of ritual and ritualized event processing. *Journal of Cognition and Culture*, *13*(3-4), 347–365.
Schjoedt, U., & Andersen, M. (2017). How does religious experience work in predictive minds? *Religion, Brain & Behavior*. doi:0.1080/2153599X.2016.1249913
Schjoedt, U., Sørensen, J., Nielbo, L. K., Xygalatas, D., Mitkidis, P., & Bulbulia, J. (2013). The resource model and the principle of predictive coding: A framework for analyzing proximate effects of ritual. *Religion, Brain & Behavior*, *3*(1), 79–86.
Taves, A., & Asprem, E. (2017). Experience as event: Event cognition and the study of (religious) experiences. *Religion, Brain & Behavior*, *7*(1), 43–62.
Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–135). New York: Oxford University Press.
Ustinova, Y. (2009). *Caves and the ancient Greek mind: Descending underground in the search for ultimate truth*. Oxford: Oxford University Press.
van Elk, M., & Aleman, A. (2017). Brain mechanisms in religion and spirituality: An integrative predictive processing framework. *Neuroscience & Biobehavioral Reviews*, *73*, 359–378.

Predictive coding in the psychological sciences of religion: on flexibility, parsimony, and comprehensiveness

Pehr Granqvist[a] and Frances Nkara[b]

[a]Department of Psychology, Stockholm University, Stockholm, Sweden; [b]Department of Psychology, UC Berkeley, Berkeley, USA

Andersen's article is a valuable contribution to the psychological sciences of religion, and its CSR and agency detection literatures in particular. In applying a predictive coding framework (e.g., Frith, 2007) to agency detection, Andersen makes a commendable effort to replace the now-dated notion of an endogenous, domain-specific perceptual "device" that evolved to hyperactively detect agents – supernatural or not – in situations marked by ambiguous sensory input and potential danger to the individual. In its place, Andersen urges scholars to entertain the possibility that the mind detects supernatural agency in situations when prior probability for agency is high and the sensory signal ambiguous. He also argues that agency detection is only a special case of how the mind operates as a predictive machine across other domains of mental life, including other aspects of religious cognition.

CONTACT Pehr Granqvist pehr.granqvist@psychology.su.se

Andersen's contribution warrants enthusiasm, and bodes very well for future research. His extension of the predictive coding framework to agency detection adds explicit detail regarding mental processes in making inferences about supernatural agency. A major advantage with this explicit model, in contrast with the "black box" in previous agency detection theories, is that it yields testable predictions about processes in agency detection, so may prove important to genuine scientific progress in our understanding of supernatural agent inferences.

Another advantage of Andersen's predictive coding framework is its "larger emphasis on the role of cultural practices" in the generation of false positives in agency detection. Relatedly – although not elaborated further by Andersen in his article – his predictive coding framework emphasizes "the developmental capability" of cognitive systems. The sculpting influences from culture (practices, beliefs, etc.) and the developing person's relational experiences (e.g., with caregivers) have been curiously under-emphasized in earlier (largely developmentally "maturational") theories of agency detection (for a discussion, see Granqvist & Nkara, 2017).

This Bayesian approach, however, is flexible enough that it does not necessarily need to be theoretically different from evolutionary psychology. For example, a Bayesian predictive model could posit an agency module and set evolutionarily determined priors with very high probabilities, essentially rendering the system impervious to updates and environmental learning. Granted, such a model would be superfluous – a mere translation of early evolutionary psychology's agency detection system into a mathematical expression. Noting this, however, we see that what is different with Andersen's approach is not only predictive coding *per se*, but an emphasis (1) on its associated flexibility (e.g., "a stronger focus on processual and domain-general models)" and (2) on environmentally determined prior probabilities. Thus, observations previously explained by a dedicated agency detection system, and even data that were not explained, can be better modeled on the basis of prior cultural and environmental influences. As a theoretical tool, then, predictive coding might be used to model a large range of theoretical perspectives and experimental data, possibly taking into account religious developmental trajectories that are at once biologically experience-expectant *and* cultural, as well as evolutionarily ultimate *and* proximate. What is perhaps the heart of the improvement in this approach is its flexibility and potential to model integrated biological and environmental influences, as well as lifting the restriction imposed by dedicated modules that have not been corroborated by neuroscience data.

Next, we discuss critical issues, in the hopes that Andersen will address and elaborate upon them in future work. First, Andersen oversimplifies the depiction of previous models of perception (and cognition). For example, "the brain is portrayed as a passive receiver of input that only structures and organizes the input it receives in a predominantly bottom-up manner." While this charge no doubt rings true of some influential models of perception, it is not a fair characterization of how psychologists have understood perception and cognition on the whole. From expectancy effects to confirmation bias, from self-verification to self-fulfilling prophecies, from role theories to stereotype threat, psychologists have not turned a blind eye to the role of predictions based on, and biased by, past experiences and learning, when the mind interprets its present surroundings and projects future possibilities.

Indeed, going back in the history of cognitive science itself, Craik (1943; see also Young, 1964) offered relatively elaborate ideas on mental representations, with the notion that experiences build and update "internal working models" (IWMs), and also theorized these to be important for forecasting future events. Along with ethology, Craik's IWMs were a key component of attachment theory (Bowlby, e.g., 1969/1982; see also Bretherton & Munholland, 2008). Bowlby was indeed a keen reader of the cognitive science available at the time, though many cognitive and neuroscientists have misrepresented attachment theory as principally "neo-Freudian." Bowlby explicitly noted the predictive nature of IWMs: "The use to which a model in the brain is put is to transmit, store and manipulate information that helps in *making predictions* as to how what is here termed set-goals can be achieved" (1969/1982, p. 80, italics added), and "every situation we meet with in life is construed in terms of the representational models we have of the world about us and of ourselves" (1980,

p. 229). Bowlby's underlying idea was that our flexibility, adaptability, and complexity as a species are contingent on our ability to make small-scale mental predictions (based on our IWMs), to guide behavior in future situations that are to varying degrees similar to ones we have already encountered. Its strong resemblance to the principles of predictive coding is no coincidence, as these early ideas have undergone parallel developments. For example, Craik's IWM term was independently rediscovered by cognitive scientist Johnson-Laird (1983) and subsequently adopted within cognitive neuroscience (Bretherton & Munholland, 2008).

Second, Andersen briefly touches upon fear and threat as important for religious experience, but it is not yet clear how emotion or motivation may be understood within the predictive coding framework. Although he does not make strong specific claims regarding the comprehensive potential of the predictive coding framework, he repeatedly refers to its "unifying power and great … explanatory scope" for modeling how the mind works, including religious cognition and the interpretation of associated experience. The model will nonetheless be insufficient unless its specific formulations include attention, affect, and motivation, in addition to other forms of cognition in mental processes. Many religious experiences are intensely motivational and affective, such that any comprehensive theory would need to incorporate these aspects. For example, why does the mind turn religious particularly when the individual finds him- or herself in emotional turmoil, at odds with the self and with conspecifics (see also Granqvist & Kirkpatrick, 2016)?

Relatedly, some experiences become religiously interpreted not because they are predicted, but because they are unanticipated and then very difficult for the person to understand. Mystical experiences, at least those marked by dissolution of one's usual sense of self, provide one example of such highly unexpected experience (i.e., low prior probability). Also, set/setting incongruity is a well-documented facilitator of such experiences (Hood, Hill, & Spilka, 2009), again indicating that failed expectations (more than predictions being carried forward) play a major role. These more radical violations of prediction, which can motivate religious interpretations, would also need to find their place in a comprehensive model based on predictive coding.

Finally, in our view, the major strength of the predictive coding framework is flexibility as well as parsimony in using prediction revision as a core organizing principle. Andersen has eloquently illustrated how such a parsimonious framework may be put to use in the psychological sciences of religion, especially to allow proximate cultural and other experiential influences. However, in providing an alternative model to overly fixed (evolutionarily ultimate) accounts, it runs the risk of reacting against them in ways that are not beneficial. While in some places Andersen marks the distinction between the model and actual brains, in other places the model is conflated with the brain, with the theory supplanting any account of actual anatomical, physiological, and genetic functions. What is needed is an integrative model, inclusive not only of many mental modalities and environmental-cultural influences in development, but also priors that incorporate the constraints and affordances of actual biology. We foresee that a robust theory will incorporate information on these aspects as well.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The writing of this commentary was facilitated by a grant from the John Templeton Foundation [grant number 51897].

References

Bowlby, J. (1969/1982). *Attachment and loss: Vol. 1. Attachment*. New York: Basic Books.
Bowlby, J. (1980). *Attachment and loss: Vol. 3. Loss*. New York: Basic Books.

Bretherton, I., & Munholland, K. A. (2008). Internal working models in attachment relationships: Elaborating a central construct in attachment theory. In J. Cassidy & P. R. Shaver (Eds.), *Handbook of attachment: Theory, research, and clinical applications* (pp. 102–127, 2nd ed., Chapter 5). New York, NY: Guilford Press.
Craik, K. (1943). *The nature of explanation*. London: Cambridge University Press.
Frith, C. (2007). *Making up the mind: How the brain creates our mental world*. Malden, MA: Blackwell.
Granqvist, P., & Kirkpatrick, L. A. (2016). Attachment and religious representations and behavior. In J. Cassidy & P. R. Shaver (Eds.), *Handbook of attachment: Theory, research, and clinical applications* (pp. 856–878, 3rd ed.). New York, NY: Guilford.
Granqvist, P., & Nkara, F. (2017). Nature meets nurture in religious and spiritual development. *British Journal of Developmental Psychology*, *36*, 142–155. doi:10.1111/bjdp.12170
Hood Jr., R. W., Hill, P. C., & Spilka, B. (2009). *The psychology of religion: An empirical approach* (4th ed.). NYC, NY: Guilford.
Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.
Young, J. Z. (1964). *A model of the brain*. London: Oxford University Press.

Prediction and feedback may constrain but do not stop anthropomorphism

Stewart Guthrie

Anthropology, Fordham University, Boulder, CO, USA

I thank Andersen not only for his agreeable view of the place of Guthrie (1980) in CSR and for highlighting some parts of my argument, but also for his general clarity and his provocative and timely thesis. I agree with him that including predictive coding – or prediction error minimization – in our model of the mind is likely to help us understand how, why, and within what limits we detect agents, including ones who do not exist. Certainly, Andersen is right to note that the generation and testing of hypotheses, vital to our ability to act in the world, is central to the brain and cognition.

I believe also (whereas Andersen seems less sanguine) that a predictive coding model is compatible with existing theses on HADD and anthropomorphism. Indeed, our understandings of the latter two must fit with predictive coding to account for their universal presence, despite error minimization, in prediction. (One such understanding, for example, is that models of persons are so relevant, and their justified application to phenomena hence so valuable, that even frequent experiences of having applied them mistakenly do not suppress them.) A predictive coding approach also fits views of perception as hypothesis and of the perceived world as a construction (cf. Arnheim, 1969, 1974; Frith, 2007; Guthrie, 1980, pp. 9, 10, 12; Howhy, 2013; Nietzsche, 1966).

I rather demur at Andersen's extensive use of "supernatural" and, to a lesser degree, "paranormal" without defining them. As in much writing in CSR, "supernatural" here is important, occurring some 50 times, as an ostensible aspect of religion. Its meaning, however, is uncertain. Numbers of scholars, from Durkheim (1915/1965) to Saler (1977, 1993, n.d.) and Lampe (2003), have criticized it as culture-bound and vague. Are Martians, yetis, and mermaids, for example, supernatural or merely imaginary? In much writing, the term seems to mean little more than the latter. Some advocates of the term "supernatural" specify invisibility or intangibility of entities as criteria for membership but, as Horton (1993) argued, these criteria might make atomic particles and forces supernatural too.

Others (Lohmann, 2003 is a lucid exponent) define supernatural concepts as those that attribute volition or mind to things or events that do not have them, a position that now has support from

CONTACT Stewart Guthrie guthrie@fordham.edu

studies of mind–body dualism. In such dualism, which seems a human universal, minds have priority and independence. They may be disembodied, for example, or embodied in non-human form. Concepts of gods thus have precursors in concepts of human minds (but see Hodge & Sousa, in press). In any case, "supernatural," often used to help define another indefinite and culture-bound term, "religion," deserves definitional attention.

Another quibble is that Andersen, lumping HADD with the more complex account of anthropomorphism (Guthrie, 1980, 1993) that was one of HADD's prime sources (Barrett, 2000, p. 31), writes that both accounts make notions of hard-wiring and modularity "fundamental." Yet although modularity may be relatively prominent in HADD, as its term "device" suggests, HADD's underlying argument could be made as well in domain-general terms. Some treatments of anthropomorphism, including my own (1993), do cite certain aspects of human cognition, such as the preferential attention newborns give to even rudimentary representations of faces, that appear modular. Most scholars of anthropomorphism, however, including classical ones (e.g., Spinoza, Hume, and Nietzsche) rightly describe the phenomenon as a whole as diverse, pervasive, and highly general – in short, as the very opposite of modular. Correspondingly, the cognitive neuroscience of person perception, for example, employs notions not of modules but of brain areas and of systems spanning a number of areas (Farah & Heberlein, 2007; Schilbach, Eickhoff, Rotarska-Jagiela, Fink, & Vogeley, 2008).

A related quibble is that Andersen glosses HADD's term "hypersensitive" as "overly sensitive" and writes that the claim that we are "oversensitive" to agents is fundamental to both the HADD and the anthropomorphism accounts. This is a possible but not necessary construal of "hypersensitive," a term sometimes criticized for ambiguity. More important, the notion of over-sensitivity is absent from most standard accounts of anthropomorphism. Rather, such accounts hold that our sensitivity to agents and agency is about right, now as in the past.

Last, I think Andersen has insufficiently demonstrated a central premise, that experimental studies have failed to support current theories of agent detection. (This failure is what he says must be remedied by predictive coding.) Specifically, Andersen thinks that little empirical evidence exists for the claim that agency is privileged as a default in human cognition. Here he relies on four papers. None of these, however, aimed to test that claim. Instead, they tested several issues tangential to it: the relative susceptibility of "paranormal believers" and skeptics to illusory agent detection (van Elk, 2013); the same relative susceptibility as above, but adding religious believers (Riekki, Lindeman, Aleneff, Halme, & Nuortimo, 2013); whether paranormal and conspiracy beliefs are associated with perceptual biases (van Elk, 2015); and whether concepts of supernatural agents facilitate illusory face perception (van Elk, Rutjens, van der Pligt, & van Harrevel, 2016). Their findings, as Andersen reports, are mixed. He concludes that the studies do not support current models of the HADD. This seems true; but neither do they seem to undermine those models.

Moreover, evidence that agency detection *is* a perceptual default, in humans and in other animals, is massive and interdisciplinary. Sources include anthropology, art history, ethology, cognitive neuroscience, linguistics, literary criticism, philosophy, and psychology (Guthrie, 1993, 2002, 2015, in press). They include ample experimentation. Andersen, however, concluding that detecting agents "does not have a privileged place in our perceptual system," returns us to square one: "Then how do we explain the fact that supernatural agents are so cross-culturally prevalent?" His answer abandons cognitive science for cultural anthropology: beliefs in such agents are prevalent because "people have … adopted these beliefs from other humans in their society." This leads us, however, to an infinite regress: why did those other humans hold the beliefs?

Although hypothesis testing, including predictive coding, regulates our readiness to find agency and agents including persons everywhere, it by no means eliminates that readiness, which according to much evidence is intuitive and evolved. The readiness is also, as noted, shared by non-human animals (Darwin, 1871; Foster & Kokko, 2008; Guthrie, 2002; Harrod, 2011, 2014; von Uexküll, 1934/1992), and it has an apparent selective advantage now widely cited in CSR: that a bet that some phenomenon *is* an agent offers maximal gains and minimal losses. Andersen has served CSR well,

nonetheless, by calling our attention to the brain's necessarily constant check on its own, inevitably uncertain perceptual bets, and to how that check is accomplished.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Arnheim, R. (1969). *Visual thinking*. Berkeley: University of California Press.

Arnheim, R. (1974). *Art and visual perception: The new version*. Berkeley: University of California Press.

Barrett, J. L. (2000). Exploring the natural foundations of religion. *Trends in Cognitive Sciences*, *4*, 29–34.

Darwin, C. (1871). *The descent of man, and selection in relation to sex*. London: Murray.

Durkheim, E. (1915/1965). *The elementary forms of the religious life*. New York, NY: Free Press.

Farah, M. J., & Heberlein, A. S. (2007). Personhood and neuroscience: Naturalizing or nihilating? *The American Journal of Bioethics*, *7*, 37–48.

Foster, K., & Kokko, H. (2008). The evolution of superstitious and superstition-like behavior. *Proceedings of the Royal Society B*, *276*, 1–10. doi:10.1098/rspb.2008.0981

Frith, C. (2007). *Making up the mind: How the brain creates our mental world*. Malden, MA: Blackwell.

Guthrie, S. (1980). A cognitive theory of religion. *Current Anthropology*, *21*, 181–203.

Guthrie, S. (1993). *Faces in the clouds: A new theory of religion*. New York, NY: Oxford University Press.

Guthrie, S. (2002). Animal animism. In I. Pyysiainen & V. Anttonen (Eds.), *Current approaches in the cognitive science of religion* (pp. 38–67). London: Continuum.

Guthrie, S. E. (2015). Religion and art: A cognitive and evolutionary approach. *Journal for the Study of Religion, Nature and Culture*, *9*(3), 283–311.

Guthrie, S. (in press). Religion as anthropomorphism. In J. Liddle & T. Shackelford (Eds.), *The Oxford handbook of evolutionary psychology and religion*. Oxford: Oxford University Press.

Harrod, J. B. (2011). A trans-species definition of religion. *Journal for the Study of Religion, Nature and Culture*, *5*(3), 327–353.

Harrod, J. B. (2014). The case for chimpanzee religion. *Journal for the Study of Religion, Nature and Culture*, *8*, 8–45.

Hodge, K. M., & Sousa, P. (in press). Dualism, disembodiment and the divine: Supernatural agent representations in the study of religion. In A. K. Petersen, L. Martin, J. S. Jensen, I. Gilhus, & J. Sørensen (Eds.), *A new synthesis: Cognition, evolution, and history in the study of religion*. Method & Theory in the Study of Religion Supplement Series. Leiden: Brill.

Horton, R. (1993). *Patterns of thought in Africa and the West: Essays on magic, religion and science*. Cambridge: Cambridge University Press.

Howhy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Lampe, F. (2003). Creating a second-storey woman: Introduced delineation between natural and supernatural in Melanesia. *Anthropological Forum*, *13*(2), 167–174.

Lohmann, R. I. (2003). The supernatural is everywhere: Defining qualities of religion in Melanesia and beyond. *Anthropological Forum*, *13*, 175–185.

Nietzsche, F. (1966). *Werke in Drei Bänden* (Vol. 3). Munich: Carl Hanser.

Riekki, T., Lindeman, M., Aleneff, M., Halme, A., & Nuortimo, A. (2013). Paranormal and religious believers are more prone to illusory face perception than skeptics and non-believers. *Applied Cognitive Psychology*, *27*, 150–155.

Saler, B. (1977). Supernatural as a Western category. *Ethos*, *5*, 31–53.

Saler, B. (1993). *Conceptualizing religion: Immanent anthropologists, transcendent natives, and unbounded categories*. Leiden: Brill.

Saler, B. (n.d.). *Observations on the construction of the supernatural in Euro-American cultures*. Unpublished manuscript.

Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R., & Vogeley, K. (2008). Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. *Consciousness and Cognition*, *17*, 457–467.

van Elk, M. (2013). Paranormal believers are more prone to illusory agency detection than skeptics. *Consciousness and Cognition*, *22*, 1041–1046.

van Elk, M. (2015). Perceptual biases in relation to paranormal and conspiracy beliefs. *PlosOne*, *10*(6), e0130422. doi:10.1371/journal.pone.0130422

van Elk, M., Rutjens, B. T., van der Pligt, J., & van Harrevel, F. (2016). Priming of supernatural agent concepts and agency detection. *Religion, Brain & Behavior*, *6*(1), 4–33.

von Uexküll, J. (1934/1992). A stroll through the worlds of animals and men: A picture book of invisible worlds. *Semiotica*, *89*, 319–391.

Evolved priors for agent detection

David L. R. Maij and Michiel van Elk

Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

We compliment Andersen on an outstanding theoretical article integrating agency detection experiences within the framework of predictive coding. We pursue a similar line of thinking in our work as we believe the idea is generally spot on and adequately outlines the trajectory the cognitive science of religion should follow (Maij & van Elk, 2017; Maij, van Schie, & van Elk, 2017; van Elk & Aleman, 2017; van Elk & Wagenmakers, 2017; van Elk & Zwaan, 2017; van Leeuwen & van Elk, 2017). Nevertheless, Andersen's theoretical idea can be further elaborated in at least two respects.

First, the predictive coding framework proposed by Andersen places a strong emphasis on prior expectations formed by interaction with the environment (e.g., instruction, cultural transmission, learning, and reliance on source credibility). At the same time, we should acknowledge the possibility of "evolved priors" that were selected for through a process of natural selection (Friston, Thornton, & Clark, 2012). The literature on preparedness for learning shows that organisms are prepared to learn readily about phenomena that were relevant in an evolutionary past (for reviews, see Mallan, Lipp, & Cochrane, 2013; Öhman, 2009; Öhman & Mineka, 2001; Seligman, 1971). For example, it is easier to condition people to fear animals, thunder, heights, and social events than to condition fear responses to modern threats such as cars or handguns (Öhman, 2009). Such biases are hard to explain without assuming evolved priors that predispose humans for learning specific associations. In addition, without assuming evolved priors, the "dark room problem" lurks – a philosophical argument proposed against the predictive coding framework. This argument entails that an energy-minimizing and prediction error-minimizing biological agent situated in a dark room would be unmotivated to move, as moving out of the room would increase surprise (Friston et al., 2012; Klein, 2016). Evolved priors solve this problem by defining what states are considered "surprising." When a prior model expects a light environment, the agent will immediately try to leave the room (Friston et al., 2012).

Applying the idea of evolved prior models to agency detection means that, as a result of evolutionary pressures, specific innate models have evolved that are dedicated towards detecting predator and prey. For example, babies quickly associate snakes with fear (DeLoache & LoBue, 2009) and look longer at pictures of spiders than at reconfigured and distorted images of these spiders (Rakison & Derringer, 2008). Five- and six-year-old children in cities are afraid of monsters with claws, while they are initially unafraid of urban threats (Boyer & Bergstrom, 2011; Maurer, 1965). Adults in general have a bias towards detecting threatening animal agents, as evidenced for instance by an attentional bias to prioritize emotionally threatening stimuli (Brosch & Sharma, 2005; Flykt, 2004; Lipp, 2006; Lipp, Lipp, Derakshan, Waters, & Logies, 2004). In short, these findings are what we should expect if there were to be an evolved bias (Barrett, 2005), or "prior model" as we now like to call it, for detecting agents that were behaviorally relevant in an evolutionary past. Importantly, the notion of an evolved prior model for detecting agents is in line with the HADD theory that many other scholars have proposed (Barrett, 2000; Barrett, 2005; Boyer & Bergstrom, 2011; Öhman & Mineka, 2001), although they did not frame the notion of an evolved cognitive module in terms of the predictive coding framework. Hence, the evolutionary psychological theories that Andersen refutes may be more compatible with predictive coding than currently assumed. In sum, we propose that Andersen's proposal should be extended by acknowledging the possibility

CONTACT Michiel van Elk M.vanElk@uva.nl

that evolved constraints (especially in the domain of fear and agency) exist on the potential space that priors could take.

Of course, Andersen rightly points out that several experimental paradigms have not yielded convincing evidence for a universal bias towards detecting agents, as evidenced for instance by studies on binocular rivalry (Denison, Piazza, & Silver, 2011) and the face-house categorization task (van Elk, Rutjens, Pligt, & Harreveld, 2016). However, we would like to point out that the dependent measures in these studies may have been ill suited to capture an eventual bias for agent detection, as they primarily involved the deliberate decision of whether an agent stimulus was consciously perceived. Evolved biases for agent detection might well exert behavioral effects without producing any direct accompanying reflective beliefs (McKay & Efferson, 2010) – akin to the output of the intuitions generated by System 1 (Risen, 2016). The examples discussed above also illustrate that agent-like stimuli readily trigger adaptive behavioral responses (e.g., fear conditioning) and, in many cases, we respond to potentially threatening stimuli instantaneously without deliberate perceptual decision making.

To further establish the presence (or absence) of evolved agent detection biases, we need good behavioral proxies that are more ecologically valid than the computer-based tasks used in earlier studies. Therefore, we commend Andersen's proposal to use virtual reality techniques to test their model, and we propose to infuse such studies with relevant physiological or behavioral measures indicative of (implicit) agent detection (e.g., skin conductance, approach-avoidance measures, etc.). When adopting this method in a series of virtual reality experiments that we are currently conducting, we already observed that participants often detect agents when they are objectively not present (Maij & van Elk, in preparation). This was especially the case in threatening environments, when – according to HADD logic – humans should indeed display a bias towards falsely perceiving other agents.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Barrett, H. C. (2005). Adaptations to predators and prey. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 200–223). Hoboken, NJ: Wiley.

Barrett, J. L. (2000). Exploring the natural foundations of religion. *Trends in Cognitive Sciences*, *4*(1), 29–34.

Boyer, P., & Bergstrom, B. (2011). Threat-detection in child development: An evolutionary perspective. *Neuroscience & Biobehavioral Reviews*, *35*(4), 1034–1041.

Brosch, T., & Sharma, D. (2005). The role of fear-relevant stimuli in visual search: A comparison of phylogenetic and ontogenetic stimuli. *Emotion*, *5*, 360–364.

DeLoache, J. S., & LoBue, V. (2009). The narrow fellow in the grass: Human infants associate snakes and fear. *Developmental Science*, *12*(1), 201–207.

Denison, R. N., Piazza, E. A., & Silver, M. A. (2011). Predictive context influences perceptual selection during binocular rivalry. *Frontiers in Human Neuroscience*, 5, 166. doi:10.3389/fnhum.2011.00166

Flykt, A. (2004). Visual search with biological threat stimuli: Accuracy, reaction times, and heart rate changes. *Emotion*, *5*, 349–353.

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*, 130.

Klein, C. (2016). What do predictive coders want? *Synthese*. Advance online publication. doi:10.1007/s11229-016-1250-6

Lipp, O. V. (2006). Of snakes and flowers: Does preferential detection of pictures of fear-relevant animals in visual search reflect on fear-relevance? *Emotion*, *6*, 296–308.

Lipp, O. V., Lipp, O. V., Derakshan, N., Waters, A. M., & Logies, S. (2004). Snakes and cats in the flower bed: Fast detection is not specific to pictures of fear-relevant animals. *Emotion*, *4*(3), 233–250.

Maij, D. L. R., & van Elk, M. (2017). *Threat-induced agency detection in a virtual reality environment*. Manuscript in preparation.

Maij, D. L. R., van Schie, H. T., & van Elk, M. (2017). The boundary conditions of the hypersensitive agency detection device: An empirical investigation of agency detection in threatening situations. Manuscript submitted for publication.
Mallan, K. M., Lipp, O. V., & Cochrane, B. (2013). Slithering snakes, angry men and out-group members: What and whom are we evolved to fear? *Cognition & Emotion*, *27*(7), 1168–1180.
Maurer, A. (1965). What children fear. *Journal of Genetic Psychology*, *106*(2), 265–277.
McKay, R., & Efferson, C. (2010). The subtleties of error management. *Evolution and Human Behavior*, *31*(5), 309–319.
Öhman, A. (2009). Of snakes and faces: An evolutionary perspective on the psychology of fear. *Scandinavian Journal of Psychology*, *50*(6), 543–552.
Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, *108*(3), 483–522.
Rakison, D. H., & Derringer, J. (2008). Do infants possess an evolved spider-detection mechanism? *Cognition*, *107*(1), 381–393.
Risen, J. L. (2016). Believing what we do not believe: Acquiescence to superstitious beliefs and other powerful intuitions. *Psychological Review*, *123*(2), 182–207.
Seligman, M. (1971). "Phobias and preparedness". *Behavior Therapy*, *2*, 307–320.
van Elk, M., & Aleman, A. (2017). Brain mechanisms in religion and spirituality: An integrative predictive processing framework. *Neuroscience & Biobehavioral Reviews*, *73*, 359–378.
van Elk, M., Rutjens, B. T., Pligt, J., & Harreveld, F. (2016). Priming of supernatural agent concepts and agency detection. *Religion, Brain and Behavior*, *6*(1), 4–33.
van Elk, M., & Wagenmakers, E. J. (2017). Can the experimental study of religion be advanced using a Bayesian predictive framework? *Religion, Brain & Behavior*, *7*, 331–334.
van Elk, M., & Zwaan, R. (2017). Predictive processing and situation models: Constructing and reconstructing religious experience. *Religion, Brain & Behavior*, *7*(1), 85–87.
van Leeuwen, N., & van Elk, M. (2017). Seeking the supernatural: The interactive religious experience model. Manuscript submitted for publication.

Explaining agency detection within a domain-specific, culturally attuned model

Joni Y. Sasaki[a*] and Adam S. Cohen[b,c*]

[a]Department of Psychology, York University, Toronto, ON, Canada; [b]Department of Psychology, University of Western Ontario, London, ON, Canada; [c]The Brain and Mind Institute, University of Western Ontario, London, ON, Canada

Agents, or objects that appear to act with intentions, are readily perceived (Heider & Simmel, 1944), even when no agent is actually there, and may be characterized as "supernatural" (Barrett, 2000). In his target article, Andersen argues that a Bayesian statistical framework of predictive coding may better account for existing data and make novel predictions relevant to agency detection. We discuss a few key points in response: (1) domain specificity and agency detection; (2) sources of variability in agency detection; and (3) sources of hypotheses explaining agency detection.

Domain specificity and agency detection

Despite important similarities to the Bayesian statistical framework, domain-specific accounts are mischaracterized in the target article as uninformed by "prior expectations," thus allowing for less flexibility. In fact,

> the more that the input supports the related conceptual structure, the greater is the probability that the attribution will be consistent with the input. However, when data are ambiguous or when there are gaps in the data, we can count on our conceptual structure to assist us. (Gelman, Durgin, & Kaufman, 1995, p. 181)

CONTACT Joni Y. Sasaki jsasaki@yorku.ca
*The authors contributed equally to this work.

As these words describe a domain-specific account of animacy, something similar should be true for agency detection; perceptual input, integrated with domain-relevant conceptual representations, drives inferences about the presence of an agent. Domain-specific mechanisms rely heavily on extra-perceptual information, whether it be content within the mechanisms of interest or "top-down" content delivered from other mechanisms. In fact, domain-specific accounts should precisely state what extra-perceptual information is accessible to a mechanism that enables it to operate flexibly, generating a wider array of representations or behaviors given different contexts (Barrett & Kurzban, 2006). Examples from cognitive science abound: prior events update time representations (Balsam & Gallistel, 2009), and specific contextual factors moderate social exchange computation (Cosmides, Barrett, & Tooby, 2010).

Counter to Andersen's argument, the case of culture demonstrates how prior expectations may actually rely on domain specificity. Andersen argues that predictive coding would better account for cultural variability in supernatural agency detection because it incorporates prior expectations that a person would have learned from their cultural or social surroundings, including "religious upbringing, ritual attendance, religious schooling, societal levels of religiosity, and so forth." Yet the problem is that cultural variability does not necessitate domain generality.

Take, for example, social attention. Humans show interest in what other people are attending to. Because social attention is early and reliably developing, some may assume that the social attention system is impenetrable to top-down expectations, including cultural knowledge. However, our research demonstrates that the social attention system may be *culturally penetrable*, or sensitive to cultural inputs, given that European Americans and East Asians execute different patterns of attention to social cues automatically (Cohen, Sasaki, German, and Kim, 2017). Crucially, the social attention system has domain-specific features that seem to be attuned to particular social or cultural cues in the environment. Indeed, culture often calibrates domain-specific mechanisms in order to exert its influence.

Sources of variability in agency detection

Another important consideration is that the sources of individual differences in agency detection sensitivity are due not only to cultural, situational, or otherwise "environmental" differences, but also to genetic differences. Because the Bayesian statistical framework focuses on prior experience as one of the most important sources of variability in supernatural agency detection ("pre-existing beliefs can now additionally, and perhaps more importantly, be viewed as independent drivers" of perceived experiences with supernatural agents), it may ignore the possibility that there are biologically based sources of variation and, furthermore, that these biological susceptibilities often interact with inputs from the environment, as in gene–environment interactions (G × E) (Sasaki et al., 2013; Sasaki, Mojaverian, & Kim, 2015).

For instance, exposing people to supernatural thoughts (e.g., God, divine) increases prosocial behavior, but only for those with certain genotypes of the dopamine receptor gene DRD4, who are genetically susceptible to reward sensitivity (Sasaki et al., 2013). According to this G × E research, the source of variation cannot be predicted from cultural factors alone but must instead be understood in complex interaction with biological factors.

Sources of hypotheses explaining agency detection

One final concern with the Bayesian statistical framework centers on an induction problem: for any stimulus, there is an infinite set of potential hypotheses or interpretations that are consistent with the stimulus (Chomsky, 1980; Quine, 1960). There are, in principle, an unlimited number of inferences that flow from object detection besides those about agency: the object is kin; the object is solid; the object is worth less than $10; the object weighs more than me. Why, then, does it occur to some part of the cognitive architecture that the object might be an agent? The reason may be that domain-specific mechanisms for agency detection privilege certain hypotheses when encountering particular perceptual inputs. A problem with predictive coding as described in the target article is that it

offloads the problem of where hypotheses entering into the Bayesian inference come from in the first place. We strongly agree with the claim that the mind is "a pro-active, hypothesis-generating machine that is constantly testing its hypotheses against the incoming signal," but without domain-specific mechanisms generating a set of hypotheses relevant to agency detection, a domain-general process would have to entertain a large if not infinite set of hypotheses. Mechanisms without domain-specific content to actively structure the inference cannot solve these types of problems, let alone arrive at a solution rapidly, as agency detectors regularly do.

Concluding remarks

Prior experiences or cultural factors must be incorporated in accounts of agency detection, but a revised model still needs to explain why social knowledge would be prioritized over individual learning when it comes to supernatural agency detection. A truly Bayesian model should plateau at an accurate explanation of agency detection – that is, that there are no supernatural agents given that the system should learn from having never seen a supernatural being. That the system is resistant to data from individual learning suggests some amount of domain-specific knowledge – a prior – that facilitates agency attribution, with bounded updating from socio-cultural experiences.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Balsam, P. D., & Gallistel, C. R. (2009). Temporal maps and informativeness in associative learning. *Trends in Neurosciences*, *32*, 73–78.
Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: Framing the debate.
Barrett, J. (2000). Exploring the natural foundations of religion. *Trends in Cognitive Sciences*, *4*(1), 29–34.
Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, *3*, 1–15.
Cohen, A. S., Sasaki, J. Y., German, T. C., & Kim, H. S. (2017). Automatic mechanisms for social attention are culturally penetrable. *Cognitive Science*, *41*, 242–258. doi:10.1111/cogs.12329
Cosmides, L., Barrett, H. C., & Tooby, J. (2010). Adaptive specializations, social exchange, and the evolution of human intelligence. *Proceedings of the National Academy of Sciences*, *107*, 9007–9014.
Gelman, R., Durgin, F., & Kaufman, L. (1995). Distinguishing between animates and inanimates: Not by motion alone. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 150–184). Oxford: Oxford University Press, Clarendon Press.
Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57* (2), 243–259. doi:10.2307/1416950
Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.
Sasaki, J. Y., Kim, H. S., Mojaverian, T., Kelley, L. D., Park, I., & Janušonis, S. (2013). Religion priming differentially increases prosocial behavior among variants of the dopamine D4 receptor (DRD4) gene. *Social Cognitive and Affective Neuroscience*, *8*, 209–215.
Sasaki, J. Y., Mojaverian, T., & Kim, H. S. (2015). Religion priming and an oxytocin receptor gene (OXTR) polymorphism interact to affect self-control in a social context. *Development and Psychopathology*, *27*, 97–109.

Agency detection is unnecessary in the explanation of religious belief

Aiyana K. Willard

Institute of Cognitive & Evolutionary Anthropology, University of Oxford, Oxford, UK

CONTACT Aiyana K. Willard aiyana.willard@anthro.ox.ac.uk

A bump in the night can make us fear that a burglar is in the house. This is likely contingent on our expectations that a burglar *could* be present. If that bump is heard deep in the Canadian wilderness, we are more likely to think – and are better served by thinking – that the sound is a bear. If we are skiing across snowy mountain peaks, an unexpected noise should make us fear an avalanche rather than any type of agent. Andersen's article makes a clear case for predictive coding as an explanation for when and why we detect non-present agents. These predictions are based in our beliefs about the likelihood that an agent *is* present, derived from previous experience or prior learning, rather than an evolved tendency specifically to over-detect agents. Predictive coding and Bayesian learning are a much more parsimonious explanation than hyperactive agency detection and reflect a more current understanding of how the mind works. At the same time, making this move creates another question: if we infer non-present agents in the same way we infer anything else, do we need a specific explanation of agency detection as part of the explanation of religion at all?

Bayesian inference has overtaken several areas of psychology (Hohwy, 2013), particularly in development (Gopnik & Tenenbaum, 2007; Gopnik & Wellman, 2012), and more recently in explaining how we think about minds (Barrett, 2015). It functions as a general learning process and offers a powerful explanation for how we flexibly interact with and learn about the world. If this type of general process is to replace the hyperactive agency detection device (HADD) – a move I support – the argument becomes one in which people use their existing knowledge to simulate plausible explanations for ambiguous experiences and thus support their beliefs. If a person has been taught that rustling in the bushes is caused by ghosts, they will interpret a rustling as being a ghost and reaffirm their existing beliefs (see Barrett & Lanman, 2008). This means that once supernatural agent beliefs exist, they are maintained in the same way any other belief or experience is maintained. This is a reasonable claim, but it makes a specific role for agency detection in producing or maintaining religious belief unnecessary.

The two paths Andersen takes to explain the role of the new predictive detection of agency in explaining religion – minimally counterintuitive (MCI) memory bias and ritual practice – do not solve this problem. MCI theory has been previously used to explain the relationship between HADD and religious belief by suggesting that MCI content creates the agents that HADD reconfirms (Barrett & Lanman, 2008). Still, MCI theory suffers from similar issues to HADD; it cannot explain how we get from a low-level process to religious belief. Even if MCI concepts show a persistent memory bias and religious concepts are frequently MCI (both points have been contested elsewhere; see Purzycki & Willard, 2016), we still need to explain how increased memorability creates the belief in supernatural agents rather than just more memorable stories (see Atran, 2002; Gervais & Henrich, 2010; Willard, Henrich, & Norenzayan, 2016). A better case has been made for ritual as playing an important role in the maintenance and spread of supernatural beliefs (Henrich, 2009; Lanman, 2012; Lanman & Buhrmester, 2016; Willard & Cingl, 2017). Still, there is no clear evidence that rituals create the beliefs they support and they cannot explain the preference for supernatural agents across cultures.

By changing the underlying mechanism from an evolved module to predictive coding, we remove anything that makes this theory special to religion. HADD was meant to explain why supernatural *agents* exist so widely across cultures (Barrett, 2000). Without a preference for detecting agents, the ability to reaffirm an existing belief from ambiguous cues cannot explain the preference for supernatural *agents* over other possible beliefs. This removes the relevance of "detecting agents" in explaining religion and replaces it with a standard confirmation bias. The only thing that separates the false detection of agents from the false detection of avalanches is our culturally learned expectations.

The current best explanation for why supernatural agents are so widely present in religion is that they are inferentially rich (Boyer, 2001). We give things minds to explain unexplained phenomena because mental states are compelling explanations that help us feel like we have some ability to effect change in the world (Epley, 2014; Waytz et al., 2010). The process of reasoning about something's mind (real or imagined) is a slow, effortful, and motivated process rather than an automatic response to diminished cues (Apperly & Butterfill, 2009). Further, it is not clear that we give all – or even most –

of what we perceive as agents full-blown minds (Apperly, Riggs, Simpson, Chiavarino, & Samson, 2006; Butterfill & Apperly, 2011). Ghosts and gods are given thoughts, desires, and beliefs. We reason about them as minds, not just agents that make a bump in the night. This ability to explain non-mentalistic phenomena with human-like mental states may be the more important piece of the puzzle.

A clear case has been made for Bayesian inference as the source of our mistaken detection of agents, but by making this move, Andersen makes a strong case for abandoning agency detection as a relevant part of explaining religion altogether. Its new role in religion can be summed up in a single sentence: religion changes what people expect out of the world, and therefore, based on general systems of expectation and prediction, leads people to explain their experiences in terms of their religious beliefs. Incorrectly detecting agents from ambiguous stimuli may offer us nothing specific in predicting or reinforcing supernatural beliefs. It may be time to drop the idea of agency detection altogether and start looking at the much harder problem of why humans are motivated to use mental states to predict the behavior of things that do not have minds.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970. doi:10.1037/a0016923

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, *17*(10), 841–844. doi:10.1111/j.1467-9280.2006.01791.x

Atran, S. (2002). *In gods we trust: The evolutionary landscape of religion*. Oxford: Oxford University Press.

Barrett, H. C. (2015). *The shape of thought: How mental adaptations evolve*. New York: Oxford University Press.

Barrett, J. L. (2000). Exploring the natural foundations of religion. *Trends in Cognitive Science*, *4*, 29–34.

Barrett, J. L., & Lanman, J. A. (2008). The science of religious beliefs. *Religion, Brain & Behavior*, *38*, 109–124.

Boyer, P. (2001). *Religion explained: The evolutionary origins of religious thought*. New York: Basic Books.

Butterfill, S. A., & Apperly, I. A. (2011). How to construct a minimal theory of mind. *Mind and Language*, *28*, 606–637.

Epley, N. (2014). *Mindwise: How we understand what others think, believe, feel, and want*. New York: Random House.

Gervais, W. M., & Henrich, J. (2010). The Zeus problem: Why representational content biases cannot explain faith in gods. *Journal of Cognition and Culture*, *10*, 383–389.

Gopnik, A., & Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive development. *Developmental Science*, *10*(3), 281–287. doi:10.1111/j.1467-7687.2007.00584.x

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108. doi:10.1037/a0028044

Henrich, J. (2009). The evolution of costly displays, cooperation, and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior*, *30*(244-260), 244–260. doi:10.1016/j.evolhumbehav.2009.03.005

Hohwy, J. (2013). *The predictive mind*. New York: Oxford University Press.

Lanman, J. A. (2012). The importance of religious displays for belief acquisition and secularization. *Journal of Contemporary Religion*, *27*(1), 49–65. doi:10.1080/13537903.2012.642726

Lanman, J. A., & Buhrmester, M. D. (2016). Religious actions speak louder than words: Exposure to credibility-enhancing displays predicts theism. *Religion, Brain & Behavior*, 1–14. doi:10.1080/2153599X.2015.1117011

Purzycki, B. G., & Willard, A. K. (2016). MCI theory: A critical discussion. *Religion, Brain & Behavior*, *6*(3), 207–248. doi:10.1080/2153599X.2015.1024915

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3), 410–435. doi:10.1037/a0020240

Willard, A. K., & Cingl, L. (2017). Testing theories of secularization and religious belief in the Czech Republic and Slovakia. *Evolution and Human Behavior*, *38*(5), 604–615. doi:10.1016/j.evolhumbehav.2017.01.002

Willard, A. K., Henrich, J., & Norenzayan, A. (2016). Memory and belief in the transmission of counterintuitive content. *Human Nature*, 1–23. doi:10.1007/s12110-016-9259-6

RESPONSE

The Bayesian observer and supernatural agents

Marc Andersen[a,b,c]

[a]Department of Culture and Society, Aarhus University, Aarhus, Denmark; [b]Religion, Cognition and Culture, Aarhus University, Aarhus, Denmark; [c]Interacting Minds Centre, Aarhus University, Aarhus, Denmark

I want to thank the commentators for their delightfully constructive contributions. I am happy to see that most of the commentators welcome the application of the predictive coding (PC) framework on agency detection. The commentators point to four overall issues that would benefit from further elaboration. Thematically, these issues are related to (1) general issues with the model; (2) evolved priors; (3) threat and emotions; and (4) consequences for the study of religion. Below, I address these issues in turn.

General issues with the model

According to Granqvist and Nkara (and, to some extent, Sasaki and Cohen), the target article oversimplifies the depiction of previous models of perception (and cognition) by not fully recognizing the extent to which previous research has focused "on the role of predictions based on, and biased by, past experiences and learning, when the mind interprets its present surroundings and projects future possibilities." This accusation bears some truth, mainly because the target article did not cover the history of research leading up to the predictive coding framework. The roots of predictive coding go as far back as the work of the German physician Hermann von Helmholtz, who discovered that the human perceptual system is a probability-based hypothesis tester that uses knowledge derived from the past (Helmholtz, 1860). During the twentieth century, Helmholtz's idea was developed in various ways and influenced by a range of different approaches, including the "New Look Psychology" (Bruner, Goodnow, & Austin, 1956), the analysis-by-synthesis school (Neisser, 1967/2014), and work in machine learning and computational neuroscience to name but a few (for the history of predictive coding, see Clark, 2016; Howhy, 2013). And indeed, while dominant models of perception for a long time have relegated predictions to the fringes of perception, Granqvist and Nkara are right when they claim that not all psychologists and cognitive scientists have turned a blind eye to the role of predictions in perception and cognition (Wiese & Metzinger, 2017). One of the defining features of predictive coding, however, is that it puts an unprecedented and extreme emphasis on the role of predictions, specifying that prior knowledge and top-down processing is "a *pervasive* feature of perception, which is not only present in cases in which the sensory input is noisy or ambiguous, but *all the time*" (Wiese & Metzinger, 2017, p. 3, emphasis in the original). This is a major shift from previous accounts of human perception, one which, as I argue in the target article, also has profound consequences for research on agency detection.

Granqvist and Nkara go on to constructively challenge a PC account of agency detection by pointing to experiences that

> become religiously interpreted not because they are predicted, but because they are unanticipated and then very difficult for the person to understand. … [R]adical violations of prediction, which can motivate religious interpretations, would also need to find their place in a comprehensive model based on predictive coding.

Although I believe that Granqvist and Nkara are essentially right about this, I do find their point to be somewhat misplaced here. In the target article, I address how false perceptions themselves, and not post hoc interpretations of perceptions, may form in human minds. Still, Granqvist and Nkara point to an issue well worth investigating by highlighting the importance of embedding

CONTACT Marc Andersen mana@cas.au.dk

unexpected experiences for post hoc interpretations in the PC framework, an issue in which there is already emerging interest (Hermans, 2015).

Sasaki and Cohen take issue with PC as a processual and domain-general model of perception as such. Instead, they stick with the traditional narrative in which the mind is viewed as a collection of computationally distinct, domain-specific modules. Although the question of modularity is often a question of perspective, PC does, from a particular viewpoint, present a challenge to a classic modular view of the mind, because it offers a philosophically elegant, mechanistic, and neurally plausible account of cognition and perception in which cognitive and perceptual architectures emerge as "profoundly unified and, in important respects, continuous" (Clark, 2013, p. 187). Along the same lines, PC convincingly accounts for the fact that there are no clear anatomical boundaries suggesting hard modular architectures in the human brain (Howhy, 2013). What emerges instead is a predictive neuronal hierarchy where the brain predicts at different spatial and temporal scales at the same time, and in which different levels of the hierarchy are predictive of each other (Clark, 2016; Friston, 2009; Friston & Kiebel, 2009; Howhy, 2013). Notably, this means that since there are "no theoretical or anatomical boundaries preventing top-down projections from high to low levels of the perceptual hierarchy" (Howhy, 2013, p. 122), we should expect cognitive penetration in the sense of seeing effects of higher-level processes on low-level perception (Lupyan, 2015). In an effort to provide evidence from their own work for a classic modular view of the mind, Sasaki and Cohen present very interesting evidence that fundamental low-level processes such as social attention can be shaped by cultural learning (Cohen, Sasaki, German, & Kim, 2017). Coincidentally, these are exactly the effects one would expect from a predictive system as specified by PC.

Sasaki and Cohen go on to argue that PC has a problem in accounting for how the brain chooses its models of the world.

> [F]or any stimulus, there is an infinite set of potential hypotheses or interpretations that are consistent with the stimulus … . Why, then, does it occur to some part of the cognitive architecture that the object might be an agent? … A problem with predictive coding as described in the target article is that it offloads the problem of where hypotheses entering into the Bayesian inference come from in the first place.

As I attempted to show in the target article, priors are initially, and to some extent ultimately, what decides the most likely hypothesis within the PC framework. Some of these priors may be innate and some of these priors may be learned. The extent to which some of these priors are also hard-wired, which seemingly is what Sasaki and Cohen mean to suggest is the case with HADD, is questionable. By analogy, it has for a long time been established that the human visual system assumes light comes from above. In other words, humans seemingly have a fundamental strong prior, which quite possibly could be innate, that light always comes from above. Chickens, it turns out, also assume that light comes from above. This assumption is hard-wired into their brains, rendering them utterly confused when placed in cages where light comes from below (Hershberger, 1970). Humans, on the other hand, quite easily override this prior and can swiftly learn to adapt to changes in environmental lighting (Adams, Graf, & Ernst, 2004). In other words, here we have an extremely strong candidate for a hard-wired prior, yet humans do not seem to have it. This, I argue, is also true with regards to the perception of agents. Agents have a fundamentally revealing quality about them that makes them very easy to learn about, namely that they simply, in contrast to everything else on earth, disobey Newton's laws of motion. This, along with the evidence I present in the target article, ultimately renders a hard-wired module for agency detection unnecessary.

Finally, Sasaki and Cohen make their last stand for the classic modular mind by arguing that

> [a] truly Bayesian model should plateau at an accurate explanation of agency detection – that is, that there are no supernatural agents given that the system should learn from having never seen a supernatural being. That the system is resistant to data from individual learning suggests some amount of domain-specific knowledge … .

In other words, Sasaki and Cohen argue that if humans had Bayesian brains, they should, at some point, cease committing perceptual mistakes. This common misunderstanding is actually covered

in the target article, but I shall happily reiterate it here. In the target article, I argue that most false positives in agency detection can be seen as the result of top-down interference in a Bayesian system generating high prior probabilities in the face of unreliable stimuli. In a nutshell, we see ghosts when it is dark and we expect them to be present. Importantly, contexts of low sensory reliability (contexts of low precision) translate into the brain putting more weight on prior expectations relative to sensory input. This is another way of saying that the system down-regulates the strength of prediction errors when the sensory input is expected to be unreliable. In simplified terms, this means that we have a very hard time learning from our perceptual mistakes when they occur in contexts of low sensory reliability, because our sensory system simply does not flag any given errors that we make. This is why the Bayesian observer continuously makes the same mistake over and over again under particular circumstances, and also the reason why the Bayesian observer may even grow more prone to committing these mistakes in the future, since prior expectation of the percept grows stronger each time.

Evolved priors

The big question raised by many of the commentators concerns the question of evolved priors (Asprem; Granqvist and Nkara; Maij and van Elk; Guthrie). These commentators point out that, theoretically, there is still room for HADD within the predictive coding framework. Framed in predictive coding terms, HADD would then simply be some form of hard-wired prior, arising over the course of evolution, which, somehow, would make humans more readily expect agents relative to other things. As Maij and van Elk also demonstrate, there are many good reasons to think that humans have developed various forms of priors over the course of evolutionary time. Human organisms, for instance, may innately expect there to be only one cause of sensory input at one place at a given point in time (Clark, 2013) or have certain evolved priors governing states such as hunger, perhaps through expected levels of glucoses.

I note here that despite an overall agreement among the commentators that, *theoretically*, there is room for HADD as an evolved prior within the PC framework, no commentators, save for Guthrie, are willing to argue, or present evidence, that this is actually likely to be the case in normal human perception (for the potential effects of threat on agency detection, see below). Guthrie maintains, as the only commentator, that "evidence that agency detection *is* a perceptual default, in humans and in other animals, is massive and interdisciplinary," citing only his own work. Over the course of his work, Guthrie has convincingly and consistently shown that humans and many animals often do make perceptual mistakes, seeing agents when none are really present (Guthrie, 1993, 2002). Guthrie, however, has yet to show that humans and animals are more prone to mistakenly perceive agents than they are to make any other kind of perceptual mistake. Yes, we sometimes mistake a boulder for a bear, but we also mistake paper for plastic, lizards for leaves, and cats for cushions.

Accordingly, experimental research has also failed to lend support to this very central prediction that follows from HADD theory. Contrary to this argument, Guthrie argues that the experimental evidence referenced in the target article does not undermine the HADD model, because these procedures did not directly aim to test the claim that agents are privileged in human perception. Unfortunately, Guthrie presents no methodological or theoretical arguments as to why that should be an issue. I maintain that if human perception, as Guthrie and proponents of HADD would have it, really followed the "better safe than sorry" mantra, we should expect a response bias towards agent categorization as a function of noise levels in paradigms such as the biological motion task and the face-house categorization task, which has not been found (van Elk, 2015; van Elk, Rutjens, Pligt, & Harrevel, 2016). The same is true for Guthrie's claim that our perception by default prioritizes agents because they are important. If that were the case, we should expect a general response bias towards agents in the categorization of stimuli in these paradigms, but again, no such bias is present in the experimental evidence (van Elk, 2015; van Elk et al., 2016).

Maij and van Elk raise a good point when they argue that the dependent measures in the referenced experiments

> may have been ill suited to capture an eventual bias for agency detection, as they primarily involved the deliberate decision of whether an agent stimulus was consciously perceived. Evolved biases for agent detection might well exert behavioral effects without producing any direct accompanying reflective beliefs.

While this is certainly true and an empirical question well worth future investigation, it is important to remember that proponents of HADD have traditionally argued that HADD is the cause of *conscious* perceptions of agents. These are the types of experiences reported by religious individuals from across the world, and it is these experiences that have been the main focus of the target article.

As it stands, there is no systematic evidence that supports the idea of HADD in instances of normal human perception. Hence, I argue that the ball is now in the other court, and although I acknowledge the theoretical possibility of HADD within a PC framework, proponents of HADD will have to present systematic evidence that humans do in fact privilege agents as a default in their perceptual inferences.

Threat and emotions

Proponents of HADD have traditionally made two major statements: (1) humans have a general bias to see agents when none are really there; and (2) this bias grows stronger in situations perceived as being threatening or dangerous. While I have dedicated parts of the target article to refuting the first claim, I have not put the second claim under critical scrutiny. Appropriately, Granqvist and Nkara note that "it is not yet clear how emotion or motivation may be understood within the predictive coding framework."

Maij and van Elk come to our aid by qualifying the possibility that "evolved constraints (especially in the domain of fear and agency) exist on the potential space that priors could take." Maij and van Elk qualify this claim mainly by reference to parts of the vast literature on "preparedness." The literature on preparedness suggests that humans more readily expect specific agents (most notably snakes and spiders) because these supposedly posed highly relevant dangers to our ancestors. While I certainly acknowledge the theoretical possibility of such evolved priors, it is important to stress that there is a big difference between saying that fear increases the global preference for agents in our sensory system (as proponents of HADD have done) and saying that *specific* dangerous agents are more swiftly perceived by humans (as the literature on preparedness suggests). While the latter suggests a handful of hard-wired priors, the first suggests a more basic shift in how the brain predicts sensory input. It is also important to note here that the literature on preparedness is debated and is not quite as clear cut as Maij and van Elk present it in their response (e.g., Brosch & Sharma, 2005; Lipp, 2006). For instance, it has been argued that many of the findings on "preparedness" in humans may in reality be experimental artifacts born out of a conflation between the fearsomeness of the stimuli and the appearance of the stimuli in evolutionary history (Brosch & Sharma, 2005).

Maij and van Elk end their commentary, however, by revealing what sounds like a tremendously interesting Virtual Reality study that seems to directly test one of the core claims that HADD proponents have made: that fear increases the likelihood of false perceptions of agents (Maij & van Elk, in preparation). While the literature on the role of emotions within the PC framework is still at a developing stage, there are quite a few hypotheses out there worth a mention. One thought-provoking hypothesis consists of a reformulation of the James-Lange theory in which emotions are nothing more than an inference based on interceptive input and the context in which this input occurs (Howhy, 2013). Another possibility is that emotions are the products of comparisons between current states and desired states. The idea here is that the brain predicts the sensory consequences of multiple motor programs likely to attain desired action effects. Emotions are then the result of this comparison, infusing the organism with action readiness to carry out the motor program predicted to have the highest likelihood of success (Ridderinkhof, 2017). In other words, "[w]e feel angry as a result of readiness to strike, and feel afraid as a result of readiness to run away … " (Bull, 1945, p. 211). A third possibility is that fear, much like anxiety, represents a shift in weighting of bottom-up

sensory input relative to top-down predictions (Cornwell, Garrido, Overstreet, Pine, & Grillon, 2017). In other words, fear may be a state in which ascending prediction errors are amplified, resulting in the optimizing of stimulus detection. Importantly, such an idea would not necessarily entail a preference for the detection of agents relative to other dangerous stimuli. Intuitively, there is some sense to this. If we, for instance, find ourselves in a fearful state because a hurricane is hurtling through our city, are we more likely to have false positives of agents or falling bricks? Future experimental studies will have to implement control conditions in order to rule out the possibility that fear increases the rate of false positives in dangerous stimuli in general and not only the false perceptions of agents.

Consequences for the study of religion

Finally, in her thought-provoking response, Willard argues that the implementation of PC may equal the abandonment of agency detection as a relevant part of explaining religion altogether: "if we infer non-present agents in the same way we infer anything else, do we need a specific explanation of agency detection as part of the explanation of religion at all?" I think Willard takes the argument too far here. While I have argued that our Bayesian brains are vulnerable to perceptual false alarms of virtually any kind when the appropriate prior probabilities are high and the precision of sensory signals is low, there is still a tremendous difference in terms of the *impact* different types of false perceptions may have on human behavior. Surely, a bump in the night inferred to be a burglar encourages a different response to a bump inferred to come from the refrigerator. Similarly, false perceptions of gods and ghosts have other types of impact than false perceptions of gloves and guitars. For this reason alone, agency detection should remain a vital part of our overall explanation of religion.

PC allows scholars of religion to make qualified predictions in addressing questions on sensory experiences of supernatural agents while maintaining a sensitivity to specific cultural practices. It gives them a model with explanatory power that extends well beyond sensory experiences of agents alone. And finally, thanks to PC, scholars of religion can now update their priors and free themselves from seeing HADDs when none are really present.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the' light-from-above' prior. *Nature Neuroscience*, *7*(10), 1057–1058.

Brosch, T., & Sharma, D. (2005). The role of fear-relevant stimuli in visual search: A comparison of phylogenetic and ontogenetic stimuli. *Emotion*, *5*(3), 360–364.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York, NY: Science Editions.

Bull, N. (1945). Towards a clarification of the concept of emotion. *Psychosomatic Medicine*, *7*, 210–214.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 1–73.

Clark, A. (2016). *Surfing uncertainty: Prediction, action and the embodied mind*. New York, NY: Oxford University Press.

Cohen, A. S., Sasaki, J. Y., German, T. C., & Kim, H. S. (2017). Automatic mechanisms for social attention are culturally penetrable. *Cognitive Science A Multidisciplinary Journal*, *41*, 242–258.

Cornwell, B. R., Garrido, M. I., Overstreet, C., Pine, D. S., & Grillon, C. (2017). The unpredictive brain under threat: A neurocomputational account of anxious hypervigilance. *Biological Psychiatry*, *82*(6), 447–454.

Friston, K. (2009). The *ree-energy principle: A rough guide to the brain? Trends in Cognitive Science*, *13*(7), 293–301.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1521), 1211–1221.

Guthrie, S. (1993). *Faces in the clouds: A New theory of religion*. New York: Oxford University Press.

Guthrie, S. (2002). Animal animism: Evolutionary roots of religious cognition. In I. Pyysiäinen & V. Anttonen (Eds.), *Current approaches in the cognitive science of religion* (pp. 38–67). London: Continuum.
Helmholtz, H. (1860/1962). *Treatise on physiological optics, Vol III*. New York: Dover Publications.
Hermans, C. A. M. (2015). Towards a theory of spiritual and religious experiences: A building block approach of the unexpected possible. *Religion, Brain and Behavior*, *37*(2), 141–167.
Hershberger, W. (1970). Attached-shadow orientation perceived as depth by chickens reared in an environment illuminated from below. *Journal of Comparative and Physiological Psychology*, *73*(3), 407–411.
Howhy, J. (2013). *The predictive mind*. New York: Oxford University Press.
Lipp, O. V. (2006). Of snakes and flowers: Does preferential detection of pictures of fear-relevant animals in visual search reflect on fear-relevance? *Emotion*, *6*(2), 296–308.
Lupyan, G. (2015). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology*, *6*(4), 547–569.
Maij & van Elk (in preparation). *Threat-induced agency detection in a virtual reality environment*.
Neisser, U. (1967/2014). *Cognitive psychology: Classic edition*. New York: Psychology Press.
Ridderinkhof, K. R. (2017). Emotion in action: A predictive processing perspective and theoretical synthesis. *Emotion Review*. Advance online publication. doi:10.1177/1754073916661765
van Elk, M. (2015). Perceptual biases in relation to paranormal and conspiracy beliefs. *PlosOne*, *10*(6), e0130422. doi:10.1371/journal.pone.0130422
van Elk, M., Rutjens, B. T., Pligt, J., & Harrevel, F. (2016). Priming of supernatural agent concepts and agency detection. *Religion, Brain and Behavior*, *6*(1), 4–33.
Wiese, W., & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In *Philosophy and predictive processing* (pp. 1–18). Frankfurt am Main: MIND Group.