The value of beliefs

Authors: Ethan S. Bromberg-Martin[1]* & Tali Sharot[2]*

Affiliations: [1]Department of Neuroscience, Washington University in St. Louis, St. Louis, MO 63110, USA.[2]Affective Brain Lab, Department of Experimental Psychology, University College London, London, WC1H 0AP, UK.

*Correspondence to: t.sharot@ucl.ac.uk neuroethan@gmail.com

Abstract

We construct our beliefs to meet two sometimes conflicting goals: forming accurate beliefs to inform our decisions, and forming desirable beliefs that we value for their own sake. Here we consider emerging neuroscience evidence on how the brain motivates itself to form particular beliefs and why it does so.

Our beliefs are fundamental parts of what makes each of us unique. They are a major cause of both harmony and discord; shared beliefs bring people together while divergent beliefs can spark revolutions. In this age of the internet and social media, the ability of beliefs to both invigorate and polarize is more apparent than ever. This raises a fundamental question: how do people arrive at their beliefs?

A traditional approach to studying beliefs is grounded on the idea that people form beliefs to build an internal model of the world for the purpose of informing their decisions to help them achieve external goals, such as gaining rewards and avoiding punishments (**Figure 1A**). People therefore should be motivated to form accurate beliefs, using all the information they encounter that is relevant to their goals. For example, if you read an article that suggests that social distancing is recommended by the Centers for Disease Control (information), you may form a belief that social distancing reduces the spread of disease, and keep your distance from other pedestrians (action selection) in order to maintain your health (goal).

This account, however, does not convey the whole story. There has been mounting evidence to suggest that beliefs are more than just tools to achieve external goals. Rather, beliefs are a source of value in and of themselves, such that people are motivated to hold particular beliefs. For example, people generally prefer to believe they are correct rather than incorrect, they prefer to believe the future is bright rather than dark, and they prefer to hold beliefs with certainty rather than uncertainty. In the most striking cases people take actions to construct their desired beliefs even at the expense of their truthfulness, such as purposely ignoring information that could contradict their preferred worldview, even when this impedes their pursuit of external goals. This suggests that the brain does not merely treat beliefs as instruments to obtain external outcomes, but as an important source of *internal outcomes* that are rewarding or punishing in their own right (**Figure 1B**). For example, believing that we are certain to be awarded a grant may give us a sense of validation, creating positive emotions that act as an internal reward. On the other hand, being uncertain about the outcome may make us anxious and act as an internal punishment.

Theories of motivated belief formation have been proposed in different fields including psychology, philosophy and economics (for a review of this literature see Sharot and Garrett, 2016). However, only recently have neuroscientific studies begun to uncover how motivation influences belief formation in the brain, with critical insight derived from studying both human and non-human animals. Here we consider *how* the brain motivates itself to form particular beliefs and *why* it does so.

Neural representation of beliefs

A *belief* is commonly defined as the acceptance that a proposition is true. Because most beliefs are about hidden states of the world that cannot be observed directly, beliefs can be held with a degree of uncertainty. This includes, for example, beliefs about unknowns that are geographically remote (e.g., the existence of aliens), temporally removed (e.g., the history of our ancestors), or are obscured by noise (a street sign obscured by fog).

Across species, a common experimental approach to study beliefs is to examine how they are formed from noisy sensory perceptions (Shadlen and Kiani, 2013). Participants are asked to judge whether a sensory stimulus is in one of two types of states (e.g., whether a cloud of moving dots has more leftward or rightward motion), and are rewarded for correct reports. Thus, in order to maximize reward, participants must learn to translate noisy, ambiguous sensory signals into beliefs about the hidden state of the world, and to make reports based on those beliefs.

In these tasks the activity of single sensory neurons is consistently predictive of upcoming behavioral reports, as one would expect since these decisions are guided by sensory input. Intriguingly, however, while neurons in early sensory areas encode the stimulus with short latencies and high fidelity, it is single neurons in higher-order areas that are typically most predictive of upcoming behavioral reports, which sometimes differ considerably from the incoming sensory input. Taken at face value, this has strong implications for how the brain forms beliefs and uses them to guide action: even when all the necessary information is present in 'lower' sensory areas, it is only after several stages of processing that the activity of single neurons in 'higher' areas begins to resemble the beliefs about the world that guide decision making.

However, there is evidence for a more nuanced view. One can interpret each neural population as representing a distinct set of beliefs that guide different types of perceptions, actions, and learning. For instance, when humans are asked to follow a moving target with their eyes, their eyes can accurately track subtle changes in its movements even when their verbal reports about its movements are inaccurate (Tavassoli and Ringach, 2010). In this case, it is as if our brains are making two decisions based on different beliefs about the world. Our rapid pursuit eye movements are based on high-fidelity visual motion signals from early sensory areas, while our verbal reports lack this specialized input but are more flexible and abstract, integrating a vast panoply of information from throughout the brain including higher order priors, motives, and goals.

In this view, the brain contains multiple beliefs about the world, which guide behavior in different ways. This is a recurring motif in neuroscience, found in spatial navigation, sensorimotor adaptation, reward seeking, and many other processes. Importantly, these different beliefs can feed into and modulate each other. Early sensory signals representing beliefs about low-level stimulus features are sent to higher-order areas to inform beliefs about abstract, hidden states of the world. Conversely, high-level belief representations of what is likely or desirable can be sent back and alter sensory processing and perception.

Motivated belief formation

This interplay of beliefs has an important consequence. Whenever we must adjudicate between multiple, potentially conflicting beliefs and desires, there is an opportunity for our motivations to put their thumb on the scales and influence belief formation. For example, in a recent study humans were shown an ambiguous image created by morphing a photo of a face with a photo of a house and were asked whether the image was mostly face or mostly house (Leong et al., 2019). When the participants had a financial motive to want to believe the image was a face, they had stronger face-specific activity in the visual stream and they were more likely to judge the image to be a face (and vice versa for houses). This could potentially occur via selective attention. That is, our motivation to believe an image represents a face may lead us to direct our gaze or attention toward specific image features, thus prioritizing different visual input, creating different perceptual experiences, and forming different beliefs. This suggests that the mere desire to observe a certain category of stimulus can increase the neural representation of that category in sensory areas and alter our beliefs.

This example illustrates an important concept about belief formation. We do not form our beliefs through a passive process of simply taking in all the information that we happen to come across in the world. Instead, belief formation is an active process: we seek out information that is relevant to our interests, goals and desires. We can do so using a range of actions including directing overt attention to specific stimuli of interest, asking questions, conducting experiments and online searches. To fully understand how the brain forms beliefs, we thus need to understand how the brain motivates itself to seek out information, and how it uses this information to update beliefs.

Belief formation involves active information selection

The motivation to seek information from the environment is shared among humans and non-human animals. This is often a search for *instrumental* information: that is, information that helps select actions to obtain external rewards and avoid harm. However, if beliefs are a source of intrinsic value, then individuals should seek information above and beyond its instrumental value.

There are at least two aspects of beliefs that individuals often treat as though they are valuable in themselves: certainty and positive valence. That is, individuals are often averse to holding uncertain beliefs (e.g., being unsure whether you will receive grant funding) and/or negative beliefs (e.g., believing funding is likely to be denied). Thus, individuals may be motivated to seek information they expect will resolve uncertainty or produce positive beliefs (Charpentier et al., 2018).

For example, many organisms including humans, monkeys, rats, and pigeons actively seek out cues in their environment that will inform them in advance about the value of uncertain future outcomes, even when they know from extensive experience that those outcomes are uncontrollable. Humans and animals will even pay a chunk of their reward in exchange for guaranteed access to information (e.g. (Blanchard et al., 2015; Charpentier et al., 2018)). This is what one would expect if information provides a subjective benefit by allowing individuals to move from a state of uncertainty (“Will my reward be big or small?”) to a state of certainty (“My reward will be big!” or “My reward will be small!”). Furthermore, in many cases individuals seek information more avidly when there is a greater likelihood the information will be desirable (“My reward will be big!”) than undesirable (“My loss will be big!”). These observations suggest that the goal in seeking information is not merely to change the *external* world, but also to change *internal* belief states (**Figure 1B**). To do so, the brain appears to instruct and motivate information seeking by tapping into the same motivational circuits that serve this role for external rewards like food and water – in essence, treating information as if it gives rise to an ‘internal reward’ of its own.

In monkeys, single neurons can integrate primary reward and information into a common currency of value (Bromberg-Martin and Hikosaka, 2009). Specifically, neural systems for reward prediction errors (RPEs), including lateral habenula and midbrain dopamine neurons, are thought to encode the difference between the reward value of the current situation and the reward value it was predicted to have (roughly speaking, “actual reward – predicted reward”). Remarkably, these neurons respond in similar ways to “more/less water than predicted” and “more/less information than predicted". This seems to be the case in humans as well, where information-related RPEs are present as BOLD signals in dopamine-rich midbrain regions and their prominent reward-related projection targets (Charpentier et al., 2018). These signals are further modulated by how likely information is to be desirable. In both monkeys and humans, variations in neural information-related signals predict differences in information seeking behavior. This suggests that the brain employs the potent reinforcing and motivational effects of the RPE-driven valuation system to instruct actions to seek both primary reward and information.

Importantly, while the brain integrates the values of information and primary reward for the purpose of decision making, it does not treat information as identical to a primary reward. Instead, the brain seems to treat information as the source of a distinct form of 'internal reward' in its own right. A key line of evidence comes from recordings in the orbitofrontal cortex (OFC). The OFC is known to have a role in encoding and updating the values of different types of rewards (such as increasing the value of water when thirsty and the value of food when hungry). Indeed, OFC neurons respond to each option in a decision-making task by encoding both the expected amount of its water and the subjective value of its information. Remarkably, OFC neurons treated *amount of water* and *information about the amount of water* as distinct entities, encoding them with different neural codes (Blanchard et al., 2015). This is fascinating because it would potentially allow the OFC to regulate the value of information based on a tailored suite of internal states, just like it does for other forms of reward; for instance, by ensuring that when we are thirsty we place high value on water, and when we are curious we place high value on information.

Indeed, a recent study revealed a cortico-basal ganglia neural network that motivates information seeking under uncertainty (White et al., 2019). This network includes neuronal subpopulations in anatomically connected regions of anterior cingulate cortex (ACC), dorsal striatum (DS), and anterior and ventral pallidum (Pal). These neurons have an *information-anticipatory signal*: their signal activates when the animal is uncertain about the size of an upcoming reward, ramps up to the moment the animal expects to receive information to resolve the uncertainty, and then shuts off once the uncertainty has been resolved. In parallel with this information-anticipatory signal, monkeys made information seeking gaze shifts toward objects associated with uncertainty and its resolution. Moment-to-moment fluctuations in the neural signal predicted future information seeking gaze shifts, while inactivating nodes of the network impaired information seeking. Thus, this information seeking appears to be motivated by a neural network that explicitly tracks uncertainty about future outcomes and anticipates opportunities to resolve the uncertainty.

Using information to update beliefs

Once information has been obtained, it should be used to alter beliefs. If individuals merely used beliefs as an instrumental tool for decision making then they would process information impartially, in order to gain the most accurate picture of the world. In reality, however, individuals often update their beliefs in accordance with the idea that beliefs in themselves give rise to additional, internal rewards and punishments. In particular, individuals are more likely to update their beliefs in response to new information that is consistent with their preferred beliefs.

First, in line with the notion that people prefer to hold high confidence beliefs, it has been shown that people become substantially more confident in their beliefs when they learn someone agrees with them, but only become slightly less confident when they learn someone disagrees with them

(Kappes et al., 2020). The pMFC, which is important for monitoring and modifying judgements in response to information, fails to track the strength of others' contradictory beliefs (Kappes et al., 2020). As a result, people remain relatively confident in their beliefs even in the face of strong disagreement.

Second, consistent with the notion that positive beliefs are more valued than negative beliefs, it has been observed that animals can interpret ambiguous stimuli in a way which supports desired beliefs (for review, see (Sharot and Garrett, 2016)). For instance, animals can be more likely to judge an ambiguous auditory or visual cue as indicating an upcoming large reward rather than a small reward or punishment. A related body of work has shown that people also learn more from information that can generate positive rather than negative beliefs (Sharot and Garrett, 2016). For instance, individuals alter beliefs to a greater extent upon receiving better-than-expected information (for example, learning that the likelihood of receiving a job offer is higher than previously thought) compared to worse-than-expected information (learning it is lower). This learning asymmetry is related to weakened neural representation of errors in response to unexpected negative information.

Interestingly, these asymmetries in updating beliefs from positive vs. negative information are absent in humans with depression (Sharot and Garrett, 2016). This suggests the asymmetry may promote mental health by producing positive beliefs that lower depression. However, this asymmetry may not be beneficial in all environments. In environments rife with threats, reduced learning from negative information could incur a severe cost by leading individuals to underestimate risks and fail to take precautionary action (such as not adequately preparing to fight a strong competitor). Remarkably, in exploring this tension, it has been shown that exposing participants to a threat-laden environment elicits a physiological stress response that increases learning from negative information, eliminating the bias in belief updating (Garrett et al., 2018). This adjustment to belief updating may be an adaptive response to threating environments, where the costs of underweighting negative information, and thus generating inaccurate positive beliefs, could be particularly high.

Why beliefs have value in themselves

We have argued the beliefs are not just instruments to achieve goals, but may become goals in themselves. In particular, individuals often prefer to hold positive beliefs and hold beliefs with high certainty. To achieve this, changes in information seeking and belief updating are motivated by tapping into the same circuits that drive primary reward-seeking. However, unlike primary rewards like food, beliefs *on their own* do not directly promote survival, so why are some beliefs coded as such?

Many scientists have argued for the adaptive benefit of motivated beliefs (these are reviewed in Sharot and Garrett, 2016). The core argument is that beliefs can motivate actions which make desirable outcomes more likely. Positive expectations, for example, may increase motivation and self-efficacy causing individuals to act with more rigor to achieve desirable goals. For example, all else being equal an individual who believes they will win contested resources is more likely to do so, as their belief will increase their motivation to fight for those resources.

Even more important, in our view, is the suggestion that motivated belief formation is adaptively tuned to suit the current environment (Garrett et al., 2018). In this view, the neural circuits that regulate belief formation may flexibly adjust the relative value of beliefs as (a) means to an end vs. (b) goals in their own right, based on the relative benefits of these two approaches. For instance, these circuits may boost our curiosity when we are faced with new and uncertain environments, where we cannot calculate the precise instrumental value of each piece of information, but we do know we need to collect a lot of it for new learning. Similarly, these circuits may make us more willing to learn from negative information in threatening environments where it is necessary for survival, while encouraging a bias toward positive and certain beliefs in rich, safe environments where these beliefs may reduce stress while having little cost to survival (Garrett et al., 2018).

Whatever the evolutionary origin, understanding the process of motivated belief formation is an increasingly central necessity in our society. We each now have access to an unprecedented and vast trove of information from which we can construct our beliefs. Some of this information is accurate, some is erroneous, and some is misinformation carefully tailored to appeal to our desires. We thus rely more than ever on our ability to seek out information, sift through it, and arrive at new beliefs that are beneficial for ourselves and for society. We hope that the emerging neuroscience of belief formation will inform the development of tools to help humans successfully navigate this information-rich era.

Declaration of Interest

The authors declare no competing interests

Figure Caption.

Figure 1. The Value of Beliefs: Traditional and Revised Frameworks. (**A**) A traditional approach to studying beliefs is grounded on the idea that the value of a belief is based solely on its ability to guide action selection to optimize external outcomes (e.g., gain food and money, avoid physical harm). Following the delivery of external outcomes, individuals update their beliefs about which rewards/punishments are available and which actions lead to better outcomes. Individuals may also select actions that are expected to lead to the delivery of sensory information that can help inform future action selection. **(B)** According to the revised framework, individuals are also motivated to optimize internal outcomes (e.g., positive emotions, a sense of validation, confidence). Internal outcomes are derived directly from the beliefs themselves. Thus, individuals are additionally motivated to select actions they expect will lead to the delivery of sensory information that will generate beliefs that lead to desired internal outcomes. Individuals may also update their beliefs in response to external outcomes in way that make desired beliefs, and thus internal outcomes, more likely.

Box 1 - Studying beliefs across species

To uncover the neural basis of beliefs, we need a way to infer an individual's beliefs and a way to relate them to neural activity. A common-sense approach to infer an individual's beliefs would be to simply *ask them* ("Do you believe X is true?"). Unfortunately, we do not have direct access to the individual's beliefs so we have no straightforward way to check whether their reports are genuine. Moreover, this approach requires the individuals to understand abstract questions about their mental states and so it can only be used in humans capable of language.

We can get around these problems by inferring beliefs from non-verbal reactions and choices. Researchers can ask individuals to perform a decision-making task in which their correct course of action (e.g., the action that will yield the most reward) depends on a hidden state of the world. For instance, if the world state is X then pressing the left button on the keyboard will give a reward, while if the world state is Y then pressing the right button will give reward. The rewards can be tailored to the individuals being studied (e.g., cash for humans and treats for animals). Thus, their actions should be a sensitive readout of whether they believe the world is in state X or Y. We can then investigate how individuals convert information about the world into beliefs, by testing how their inferred beliefs change when they are provided with different types of information. Studies in humans can directly test how this approach is working by asking participants to explicitly report their beliefs and comparing them to their inferred beliefs. Meanwhile, studies in animals can use more precise methods to measure and manipulate neuronal activity, and to test how the neuronal substrates of beliefs are conserved across species.

In individuals who lack language and the ability to participate in standard choice tasks (e.g., human infants) it is still possible to study certain beliefs based on reactions such as eye movements. Individuals typically orient to unexpected events while having more muted reactions to the same events when they are fully expected. This allows a degree of inference about an individual's beliefs about what events are likely or unlikely to occur.

To infer an individual's degree of *certainty* in their belief, we can again use either direct or indirect approaches. In humans, we can ask participants to explicitly report their level of confidence in a belief. In both human and non-human animals, we can infer confidence by allowing participants to make bets on the quality of their judgements, sometimes called *post-decision wagering* (e.g. (Shadlen and Kiani, 2013; Kappes et al., 2020)). In these tasks, an individual can maximize reward by placing large bets when their beliefs are certain ("putting their money where their mouth is") and playing it safe when their beliefs are uncertain. Confidence should also influence the tendency to update beliefs; the more confident you are of your judgement, the less likely you should be to change your judgement in the face of contrary evidence. Finally, reaction times are often related to confidence, with shorter reaction times often associated with greater confidence.

Declaration of Interests

The authors declare no competing interests.

Acknowledgments

TS is funded by a Wellcome Trust Senior Research Fellowship 214268/Z/18/Z. EBM is supported by National Institute of Mental Health Award Numbers R01MH110594 and R01MH116937 to the Monosov lab. We thank Chris Kelly, Laura Globig, Irene Cogliati Dezza, Valentina Vellani, Liron Rozenkrantz, Bastien Blain, and Ilya Monosov for providing comments on a previous version of this manuscript.

References

Blanchard, T.C., Hayden, B.Y., and Bromberg-Martin, E.S. (2015). Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. Neuron *85*, 602-614.

Bromberg-Martin, E.S., and Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. Neuron *63*, 119-126.

Charpentier, C.J., Bromberg-Martin, E.S., and Sharot, T. (2018). Valuation of knowledge and ignorance in mesolimbic reward circuitry. Proc Natl Acad Sci U S A *115*, E7255-E7264.

Garrett, N., Gonzalez-Garzon, A.M., Foulkes, L., Levita, L., and Sharot, T. (2018). Updating Beliefs under Perceived Threat. J Neurosci *38*, 7901-7911.

Kappes, A., Harvey, A.H., Lohrenz, T., Montague, P.R., and Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. Nat Neurosci *23*, 130-137.

Leong, Y.C., Hughes, B.L., Wang, Y., and Zaki, J. (2019). Neurocomputational mechanisms underlying motivated seeing. Nat Hum Behav *3*, 962-973.

Shadlen, M.N., and Kiani, R. (2013). Decision making as a window on cognition. Neuron *80*, 791-806.

Sharot, T., and Garrett, N. (2016). Forming Beliefs: Why Valence Matters. Trends Cogn Sci *20*, 25-33.

Tavassoli, A., and Ringach, D.L. (2010). When your eyes see more than you do. Curr Biol *20*, R93-94.

White, J.K., Bromberg-Martin, E.S., Heilbronner, S.R., Zhang, K., Pai, J., Haber, S.N., and Monosov, I.E. (2019). A neural network for information seeking. Nat Commun *10*, 5168.